CHILDREN'S TEMPERAMENT: CLUSTER ANALYSIS OF PARENT QUESTIONNAIRES

BY

JANET L. EATON

A DISSERTATION PRESENTED TO THE GRADUATE SCHOOL
OF THE UNIVERSITY OF FLORIDA IN
PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

UNIVERSITY OF FLORIDA

1983

Copyright 1983

by

Janet L. Eaton

ACKNOWLEDGMENTS

There are many special individuals whose generous contributions are deeply appreciated. Thanks go to Jacque Goldman, Ben Barger, Keith Berg, Randy Carter, Everette Hall, and Paul Satz, who have guided me intelligently, benevolently, patiently, and insightfully. Thanks go to my parents for encouraging my curiosity and my love of learning. Thanks go to my mother, my sisters Pam, Beth, and Martha, and my grandmother Ruth for their enduring love, support, and confidence. Thanks go to RME of OFC and to Jill and John for helping me locate the centers. Thanks go to Directors Maggie, Deb, Carla, Angela, Mary B., Robin, Ann, Fran, Judy, Nancy, and Pat who went out of their way to help me collect the data. Thanks go to all of the parents who volunteered. Thanks go to L. and the other loons on the lake who double-checked the scoring. Thanks go to my favorite Wizard Brad who poured coffee, massaged shoulders, kept the phones free, and "shazammed" the data into hard copy. Thanks go to Nancy Hurley and June Sprock who taught me how to use the UF computer facilities and to Leonard who afforded me their use. Thanks go to Randy Carter for so much help with programs and subsequent explanations. Special thanks go to Roger Blashfield

who spent hours teaching me cluster analysis. Thanks go to Jane Boesch who located every impossible-to-find article. Thanks go to my favorite innkeepers Mary Anna, Leonard, and Sara whose generous hospitality, compassion, good humor, love, and pets have seemed like miracles. Thanks, kudos, and love go to some other very dear friends who have supported me throughout this journey: Carol and Jody; Dev, Barrie, Randy, and Joanne; Frank, Michelle, and Melissa; Joan and Fred; and Adele, Jill, and CRob. Thanks go to Snoogums for much more than Korbel. Thanks go to Dr. Norman Neiberg for more than his nudzhing. Thanks go to Pam and Dennis for the pencil that wrote it all. Thanks go to Molly Harrower, Gloria Steinem, and Anne Alonso. Thanks go to Lois Rudloff without whose typing skills and cheer I would be lost.

TABLE OF CONTENTS

		Page
ACKNOWLEDGMENTS		iii
ABSTRACT		vi
CHAPTER		
I	TEMPERAMENT	1
II	TYPOLOGIES AND CLASSIFICATION	19
III	CLUSTER ANALYSIS	29
IV	CLUSTER ANALYSIS AND TEMPERAMENT	48
V	METHOD	59
	Subjects	59
	Materials	64
	Procedure	67
VI	RESULTS	75
	Findings for the Total Sample (N=200)	75
	Clustering Sample A	79
	The Nearest-Centroid Technique	87
VII	DISCUSSION	93
REFERENCE NOTES		105
REFERENCES		106
BIOGRAPHICAL SKETCH		117

Abstract of Dissertation Presented to the Graduate School of the University of Florida in Partial Fulfillment of the Requirements for the Degree of Doctor of Philosophy

CHILDREN'S TEMPERAMENT:
CLUSTER ANALYSIS OF PARENT QUESTIONNAIRES

By

Janet L. Eaton

December 1983

Chairperson: Jacquelin R. Goldman, Ph.D.
Major Department: Clinical Psychology

Cluster analysis was used with questionnaire data to assess the validity of the temperament typology developed by Thomas and Chess. Parents of children in day care (ages 3-7 years) completed McDevitt and Carey's Behavioral Style Questionnaire (BSQ). There were 200 questionnaires returned with a sufficient response percentage to be included in the sample. The BSQ yields nine category scores, one for each of Thomas and Chess's temperament dimensions, and a cluster diagnosis. A comparison of this sample and the BSQ standardization sample revealed significant differences between the groups' means for six categories and the variances for one category. Questions concerning the generality of the BSQ are raised. There were few age- and sex-effects. McIntyre and Blashfield's nearest-centroid

evaluation technique was used to assess the validity of the cluster solution. Accordingly, the sample was randomly divided into two groups, Sample A (n=99) and Sample B (n=101). Diverse clustering methods were used with Sample A, and the solutions were compared. Four clusters were fairly consistent across methods. The procedure which had produced the solution selected as best involved the removal of outliers, followed by clustering using the minimum variance method (with distance) and an iterative partitioning method to relocate any poorly-assigned individuals. Sample B was cluster analyzed using that procedure. Also, members of Sample B were assigned to the nearest centroid of Sample A. The degree of agreement between these two classifications of Sample B supported the validity (although minimally) of the empirically derived typology. The characteristics of one of the three stable clusters concur with Thomas and Chess's description of the slow to warm up child. The two remaining stable clusters both match Thomas and Chess's description of the easy child. Consideration of the additional temperament categories revealed that one of these easy groups was much less distractible, was less active, and had a higher sensory threshold. No support was found for the type known as the difficult child. These findings, possible limitations of the study due to selection bias, and implications for future work are discussed.

CHAPTER I
TEMPERAMENT

During the past 25 years, interest in the concept of temperament burgeoned among child development professionals and personality researchers. Recent investigations are quite diverse in terms of such design elements as conceptualization of temperament (e.g., Buss & Plomin, 1975; Korner, 1982; Plomin, 1982a, 1982b; Rothbart, 1982; Rothbart & Derryberry, 1981; Standley, Soule, Copans, & Klein, 1978), subject population (e.g., Cadoret, Cunningham, Loftus, & Edwards, 1975; Carey, 1970, 1972a, 1972b, 1974; Dunn & Kendrick, 1980; Gunn, Berry, & Andrews, 1981), and the purpose of the study (e.g., Beebe & Sloate, 1983; Berberian & Snyder, 1982; Billman & McDevitt, 1980; Cameron, 1977, 1978; Dreger, 1968; Keogh, 1982a; Matheny & Dolan, 1980; Persson-Blennow & McNeil, 1981). Yet despite its diversity, much of the literature has been stimulated by the pioneering contributions of one team, Thomas and Chess and their colleagues (1957, 1961, 1963, 1968, 1970, 1977, 1980, 1982). It is their theoretical framework which serves as the basis for this study as well.

Interested in individual differences in behavior, Thomas, Chess, Birch, Hertzig, and Korn (1963) initiated an extensive systematic longitudinal research project called

the New York Longitudinal Study (NYLS) in 1956, with follow-ups (Thomas & Chess, 1977, 1980; Thomas, Chess, & Birch, 1968). Its major focus was the anterospective exploration of individuality in behavioral styles, or temperament--its patterns, and stability as well as its influence on later psychological development. They use the term temperament as a general, phenomenological term that refers to the how of behavior, not the how well or the why or the what. The focus is on the style of behavior,[1] not such other possible components as content, ability, or motivation (Cattell, 1950; Thomas & Chess, 1977). Their definition makes no etiological statement. Temperament is simply considered a characteristic of the organism. Its expression and its actual nature are understood within an interactionist framework: as such, development involves a continual interplay between temperament and environmental factors.

In the early phase of the NYLS, analyses led to the identification of nine formal characteristics of behavior (where formal aspects are those present in a range of behaviors irrespective of content). Using these nine

---

[1] In much of the temperament literature, the terms temperament and behavioral style are used interchangeably. Recently, however, Thomas and Chess (1980) have proposed that "temperament be used to designate those stylistic characteristics which are evident in the early infancy period, while the broader term behavioral style be used for those characteristics which appear in later childhood or adult life" (p. 73).

categories, Thomas et al. (1968) identified three patterns of temperament, or temperament constellations, considered to be functionally significant. These were called the difficult child, the easy child, and the slow to warm up child. This study addresses the issue of the validity of these constellations. Inasmuch as the method and results of the NYLS have been published in detail (Thomas & Chess, 1977, 1980; Thomas et al., 1963, 1968) and succinctly summarized (Keogh & Pullis, 1980), only a brief review of those aspects most salient to this study is provided here.

Beginning in 1956, Thomas et al. (1963) collected data on 141 children from 85 families. (During the study, there was an attrition of five children.) The NYLS sample was mostly middle-to-upper-middle class, white, Jewish, and urban and suburban, having been selected to be fairly homogeneous in order to minimize geographical, economic, and sociocultural differences as sources of variance in individuals.

Information was collected about the children from their infancy through early adolescence, with periodic subsequent follow-ups. Much of the data was provided by the children's parents. They participated at regular intervals during which they were asked to describe not only what their child did in daily situations but also <u>how</u> he/she did it. After the child began to attend school, information was also obtained from the teacher. Direct observations were conducted

during school and during the administration of psychological tests on repeated occasions. Additional data were collected, such as measures of the child's cognitive functioning, relevant medical information, measures of child care practices and parental attitudes, and psychiatric evaluations when indicated.

Thomas et al. (1963) were interested in initially identifying a set of formal behavioral characteristics that would enable categorization of individuality. Using an inductive content analysis of the first two years' behavior protocols for the first 22 children, they were able to identify nine dimensions of formal behavioral attributes which were sufficiently wide to differentiate among individuals within each category. Although the dimensions were not considered totally independent, neither intercorrelations nor halo effects contributed significantly to rater reliability. Interrater reliability was high.

The nine temperament dimensions, employed in the NYLS and in many subsequent studies of temperament, are paraphrased (Thomas et al., 1963, 1968) as follows:

1. Activity level--the level, tempo, and frequency from which a motor component is present in the child's functioning.
2. Rhythmicity--the degree of regularity among the repetitive biological functions, including rest

and activity, sleeping and waking, eating and appetite, and bowel and bladder.

3. Approach-Withdrawal--how the child typically reacts to any new stimulus, such as food, people, toys, or procedures.
4. Adaptability--the ease with which the child's initial response pattern can be modified in the direction desired by the parents or others.
5. Intensity--the energy content of the child's response, regardless of whether that response is positive or negative.
6. Sensory threshold--the level of extrinsic stimulation necessary to evoke a discernible response.
7. Quality of mood--the amount of pleasant, joyful, friendly behavior, as contrasted with unpleasant, crying, unfriendly behavior.
8. Distractibility--the effectiveness of extraneous environmental stimuli in interfering with, or in altering the direction of the child's ongoing behavior.
9. Persistence--the child's maintenance of an activity in the face of obstacles to its continuation.

Thomas et al. have not asserted that these nine dimensions comprise a comprehensive list of characteristics of temperament, just important aspects. In fact, in their most recent book on development, Thomas and Chess (1980) discuss the likelihood of additional dimensions.

Subsequent studies have provided support of the adequacy of these nine dimensions in describing the formal characteristics of behavior, i.e., the temperament, of diverse populations of children. These populations have included the following:

-children of working-class Puerto Ricans (Hertzig, Birch, Thomas, & Mendez, 1968)

-children born prematurely who are neurologically impaired (Hertzig, 1974; Thomas & Chess, 1977)

-children who are mentally retarded (Chess & Hassibi, 1970; Chess & Korn, 1970; Thomas & Chess, 1977)

-children with congenital rubella (Chess, Korn, & Fernandez, 1971), and

-children on an Israeli kibbutz (Marcus, Thomas, & Chess, 1969).

Other investigators (Garside, Birch, Scott, Chambers, Kolvin, Tweddle, & Barber, 1975; Graham, Rutter, & George, 1973), working independently of the NYLS, yet using that conceptualization of temperament, have also identified the same behavioral characteristics in children from other countries.

The evidence for the generality of these categories of temperament begins to address the issue of construct validity (Cronbach & Meehl, 1955), in that it supports these formal characteristics as being fairly independent of culture, intellectual level, and investigator.

Findings from the initial phase of the NYLS (Thomas et al., 1963) also support the temperament construct. It was shown that children did show distinct individuality in temperament in the first weeks of life, with behavioral characteristics beginning to show a consistency of patterning after the fourth week. The behavioral characteristics were found to be independent of the parents' style of child management or their personality style. As the children developed, it was found that temperament characteristics "tended to persist" in most children over the years. Further research is necessary to partial out methodological contributions to the variability of temperament over time versus actual changes. In and of itself, change in temperament does not necessarily detract from the validity of temperament as a construct; given an interactionist framework, temperament is not viewed as either fixed or immutable. Instead, variability hints at the complexity that will be required of any sufficient model of developmental process for individuals.

As the NYLS proceeded, a subgroup came to attention before the children had reached the age of two years. Parents and diverse members of the research team described these children "in terms of a series of pejorative labels, ranging from the expression 'difficult children' by the more sedate and formal . . . to 'mother killers' by the more graphic and less inhibited" (Thomas et al., 1968,

p. 75). When the difficult children's scores on the nine temperament dimensions were compared to those for the remainder of the NYLS sample, differences were found. The difficult child was characterized in terms of the following attributes:

-irregularity of biological functions,

-predominance of negative withdrawal responses to new stimuli,

-slow adaptation,

-high frequency of expression of negative mood, and

-predominance of intense reactions.

None of these children was considered disturbed at that time. Rather, these variations fell within the boundaries of normal limits. This, then, became the first constellation of characteristics of temperament, or the first type of temperament, to draw attention.

Throughout the study, parents reported whatever problems their child was experiencing or creating. In many instances (31 children), the problems were assessed as either age-specific or as consequential to parental mismanagement. Guidance was provided to the parents of the 31 children, which resulted in a disappearance of both the parental concerns as well as the problem (Thomas et al., 1968). In some instances, providing guidance for the parents did not result in improvement. These children were then recommended for a psychiatric evaluation. (This was also the

recommendation for several children whose presenting problems were sufficient to elicit immediate evaluation.) In the first 10 years of the study, 42 children (31% of the sample) were diagnosed as having significant behavior problems. This group is referred to as the clinical sample.

By studying the relationship of temperament characteristics to behavior disorder (Thomas et al., 1968), it was observed that a sizeable proportion of the children in the clinical sample had been identified previously as difficult (70%). While only 4% of the nonclinical sample and 10% of the total NYLS sample were difficult children, the proportion jumped to 23% for the clinical sample. Taken in conjunction, these findings seemed to confirm the functional significance of the difficult temperament constellation.

Within the clinical sample, other children were conspicuous because their temperamental organization was quite opposite from the difficult children. These children were remarkably pleasant, easy children to care for. The easy child was characterized by the following:

-positiveness in mood,

-regularity in biological functions,

-low or moderate intensity of reaction,

-adaptability, and

-a positive approach to new situations.

Relative to the difficult children, and in relation to their proportion of the total sample, reported by Thomas and Chess

(1977) as 40%, easy children comprised only a small number of the clinical sample. [The exact number is not reported by Thomas et al. (1968).] As was true for the first constellation to be identified, most of the NYLS total population's easy children did not develop problems.

Thomas, Chess, and Birch (1968) performed a factor analysis of the nine temperament dimensions, the results of which were interpreted as support for these two constellations. Three factors emerged; the authors discussed only the first. That factor, Factor A, was relatively consistent for the first five years of life. Easy children, as categorized qualitatively before factor analysis, corresponded to high Factor A plus regularity; difficult children to low Factor A plus irregularity. However, other factor analyses of temperament data (Keogh, Pullis, & Cadwell, 1982; Lerner, Palermo, Spiro, & Nesselrode, 1982; Rowe & Plomin, 1977) have not consistently replicated Factor A. Methodological differences render it impossible to account for the different findings; instead, these differences underscore the necessity of further research to validate these constellations.

Another qualitative grouping led to the delineation of a third pattern (Thomas et al., 1968). The clinical sample had been divided into two groups: children with "active" symptoms (n=34) and children with "passive symptoms (n=8). "Children in the passive group were largely

nonparticipators. . . . To be included in the passive group, it was essential that the youngster show neither evidence of anxiety nor defenses against anxiety" (p. 34). Using the scores for the nine temperament dimensions, children with passive symptoms were compared to the children in the nonclinical sample (n=66), considered a control group.

The descriptive presentation of the results (Thomas et al., 1968) is confusing, seemingly inconsistent. It is reported that the children with passive symptoms "differed significantly in the magnitude of their weighted category scores from the nonclinical group only in the fourth and fifth years of life" (p. 60). Yet later in the book, it is stated that the children with passive symptoms "tended as a group in infancy and the preschool years to show low activity level, initial withdrawal responses, slow adaptability, low intensity of reactions, and a relatively high frequency of negative mood responses than did the nonclinical cases" (p. 92). Neither the t test for differences between group means nor the analysis of variance was significant for adaptability, one of the listed characteristics. Perhaps this apparent discrepancy could be clarified by examining the specific results of the analysis of variance; however, only a summary of findings is published (p. 61).

Nonetheless, it is that set of distinctive features which Thomas et al. (1968) present as characteristic of a third temperament constellation. The slow to warm up child is characterized by the following:

-low activity level,

-withdrawal from new situations and stimuli,

-slow adaptability,

-mild intensity of reactions, and

-somewhat negative in mood.

In view of the extremely small size of the derivation sample as well as its nonrandom selection, and the unclear presentation of results, it seems that further research is necessary to substantiate this constellation.

Using the three sets of distinguishing features as assignment criteria, Thomas et al. (1968) reviewed the temperament profile of every subject. Easy children comprised 40% of the total NYLS group, difficult children--10%, and slow to warm up children--15%. The three types, therefore, accounted for 65% of the sample. An investigation of the relationship between these types and behavior disorder pointed to distinctive developmental consequences for each particular pattern. As clinicians, it was not surprising to find that difficult children accounted for the largest proportion of the behavior disorder cases, with the slow to warm up children accounting for the next-largest proportion, and the easy children for the smallest. By comparison, 70% of the difficult children developed behavioral problems versus only 18% of the easy children. Although no one type was pathognomonic, it seemed that difficult children were at a much greater risk.

The conclusion that combinations of traits or constellations tended to lead to an increased risk for developing behavioral disorders was an exciting one. It seemed to point to a path toward preventative intervention. Thomas, Chess, and Birch (1968) and then, later, Thomas and Chess (1977) called for various professionals to utilize these types of temperament in their work. "Parents and teachers can now be helped to carry out their responsibilities to interact appropriately with temperament characteristics by the availability of short questionnaire forms for the infant and 3-7 year age periods which expedite the delineation of temperament" (Thomas & Chess, 1977, p. 184).

Subsequently, numerous studies have been conducted which examine the relationship between temperament and various dependent measures. In their 1977 book, Thomas and Chess have cited many of these as evidence of the validity of temperament and type of temperament. Briefly, temperament has been studied in terms of its relationship to school achievement and school adjustment (Carey, Fox, & McDevitt, 1977; Gordon & Thomas, 1967; Scholom & Schiff, 1980), parent-child interaction (Bates, Olson, Pettit, & Bayles, 1982; Cameron, 1977, 1978; Chamberlin, 1978; Simonds & Simonds, 1981; Thomas & Chess, 1977), other interpersonal relations (Thomas, Birch, Chess, & Robbins, 1961; Thomas & Chess, 1977), and behavior disorders (Cadoret et al., 1975; Graham,

Rutter, & George, 1973; Terestman, 1980; Thomas & Chess, 1977; Thomas et al., 1968).

Although conclusions are usually drawn about difficult children or easy children or slow to warm up children, the methodologies of the majority of these studies fail to investigate the variable type of child. Instead, the separate dimensions of temperament (also called the categories of temperament) are the independent variables. Typically, a few of the numerous dimensions are found to be significantly related to the dependent variable. Then, these significant dimensions are compared to the diagnostic criteria for each of the NYLS types (Thomas et al., 1968). If the two groups largely coincide, the results are interpreted as support of, even evidence for, the temperament types. Yet, clusters of traits cannot be assumed to be equivalent to clusters of individuals. The former compares variables; the latter compares individuals. It is incorrect to present interrelated variables as definitive evidence of interrelated individuals, i.e., a group of individuals distinct from others (Garside & Roth, 1978).

A few studies have been designed with temperament type as the independent variable. For example, Berberian and Snyder (1982) investigated the relationship of temperament to stranger reaction in infants. Subjects were assigned to one of five types (easy, slow to warm up, difficult, intermediate high, intermediate low) on the basis of their

temperament profiles. Slow to warm up children were dropped from inter-group comparisons due to the extremely small sample size. Easy and intermediate low children were combined to form the "easier" group of children; difficult and intermediate high children became the "harder" group. When easier children were compared with harder children, there were no significant differences in their reactions to strangers. When relating category scores to stranger reaction, 31 of 100 correlations were significant. All of these were in the expected direction, "confirming the prediction that fussy, more difficult infants tend to show more fearful and fewer friendly behaviors toward the stranger than do the more easy-going infants" (p. 84). These findings offer stronger support for the validity of temperament dimensions than of temperament type. Campbell (1979) also altered the operational criteria for temperament type. In this study of infant temperament and mother-infant interaction, difficult child was operationalized as one standard deviation above the sample mean on the rhythmicity, adaptability, and mood scales. Although the rationale for altering definitional criteria is often easily comprehended, the changes in definition and method make it difficult to compare the results of separate studies, which limits their potential usefulness. Campbell found that mothers of difficult infants tended to interact less and

respond less to their infants than mothers of the control group, offering some support of the difficult child construct.

Carey, Fox, and McDevitt (1977) conducted a study of temperament as a factor in early school adjustment. Contemporaneous measures of temperament and school adjustment showed a significant relationship between the adaptability scale and teacher judgments of school adjustment. The finding that easy children were not found to be significantly better in terms of school adjustment than other types is not consistent with theoretical predictions, underscoring the importance of further validity studies to the development of the meaning of temperament.

The little support that exists for temperament types specifically is attenuated by a consideration of several other issues. First, there have been criticisms (e.g., Bates, 1980; Persson-Blennow & McNeil, 1979; Rothbart, 1982) of several methodological limitations of the NYLS. These include the restricted nature of the subject population, the use of dimensions derived from a relatively small number (n=22) of two-year-olds' behavioral data without replication at other developmental levels, the lack of independence of the nine dimensions, and somewhat subjective and difficult-to-replicate procedure of interviewing and scoring interview information, the inconsistent methods used to identify each temperament constellation, the lack

of operational definitions for each constellation, and the extremely small sample size for the derivation groups, especially for slow to warm up children where there were eight subjects.

Secondly, a controversy has emerged in the literature concerning the reality of the concept of temperament. It has been argued that temperament data, most of which depend upon parents' reports of a child's behavior, cannot be assumed to be pure reflections of the child (e.g., Bates, 1980; Sameroff, Seifer, & Elias, 1982; Bates, Note 1). Instead, temperament is most often a measurement of a parent's perceptions, which may or may not accurately reflect a child's characteristics. Recently, the controversy has diminished, with the general consensus (Bates et al., 1982; Kagan, 1982; Lyon & Plomin, 1981; Thomas et al., 1982) supporting the validity of parents' reports. Even if temperament rating scales completed by parents are, in part, measuring parent variables in addition to characteristics of the child, that does not eliminate the potential usefulness of such information. After all, parents' perceptions of the child influence their interaction with that child. Further methodological studies are necessary in order to determine the relationship between a child's temperament and the parents' perceptions and descriptions of their child.

Given methodological and conceptual concerns, and the paucity of direct support for the three temperament constellations or types (Thomas & Chess, 1977; Thomas et al., 1966) it seems premature to recommend these types for use in any intervention program. Instead, more work concerning the validity of these types is called for.

CHAPTER II
TYPOLOGIES AND CLASSIFICATION

The process of identifying types is actually an aspect of classification. Classification, a basic process of most science (Hempel, 1966), involves the ordering of objects into groups according to their relationships, often their similarity (Bailey, 1973). Its purposes are numerous. Much of classificatory work is directed at describing "natural systems." "If it is the purpose of science to discover the true nature of things, then it is the purpose of a correct classification to describe objects in such a way that their 'true' relationships are displayed" (Sokal, 1974, p. 1116). It allows description of the structure and relationship of objects within and between groups. It aims to simplify the relationships in such a way that general statements can be made concerning groups of objects. Classifications are often heuristic, leading to testable hypotheses and theory development. In sum, classifications are necessary for communication, for information retrieval, as descriptive systems, to make predictions, and as a source of ideas (Blashfield & Draguns, 1976).

The study of classification systems can be delineated into several areas. Blashfield and Draguns (1976) describe them as follows:

> Taxonomy refers to the theoretical study of classification. Classification, in the narrow sense of the term, refers to the process of forming groups from a large set of entities or units. The term classification system is used to refer to the product of the process of classification; a classification system consists of a defined set of entities. Finally, identification refers to the process of assigning an entity to a category in an already existing classification. (pp. 140-141)

If the "entity" is a person, then the category is a type, and the system is a typology. Although there are various usages of the term, a type can be defined as a member of a category within some kind of a classification system (McQuitty, 1967) or, more specifically, as "the most representative pattern in a group of individuals located by a high relative frequency--a mode--in the distribution of persons in a multidimensional space" (Cattell & Coulter, 1966, p. 239). Typological classifications are utilized when one is interested in making generalizations about the behavior of the whole person, not about emotions, cognitions, abilities, or other features (Wood, 1969). A premise underlying the notion of types is that there are individuals who are similar and/or different along the dimensions of interest. The rationale is that the identification of types will lead to a better and fuller understanding of development through research.

Historically, much of psychology has utilized the normative model in assessing many aspects of children's

development. This approach compares homogeneous normal distributions of characteristics and linear relations between them. However, there has been an accompanying realization that more complex natural distributions need to be considered to accommodate the intricacies of human behavior. Keogh and Pullis (1980) have elaborated upon this as follows:

> It is not possible to make assumptions as to distributional characteristics of temperament dimensions. . . . It seems reasonable that an ipsative approach involving study of intraindividual organization of behavioral attributes may be a productive one for studying particular children or groups of children. The real goal of an ipsative approach is to document behavioral attributes in an idiographic, nonnormative fashion. The ipsative approach is important from a developmental standpoint in that structural changes in temperament may be viewed over time, yielding information about integration of structures or differentiation of components. (p. 271)

The magnitude of this task can be imposing, given that each person is unique and that the number of variables associated with him/her is infinite. Generalizations become more manageable as people are typed, i.e., as kinds of children are identified by a number of relevant characteristics (Wood, 1969).

Typologies vary from ideal types to empirical types (Achenbach, 1981; Skinner, 1981; Wood, 1969). Ideal types are constructs based upon theory. They are denoted by a

hypothetical pattern of attributes which is characteristic of a subset of individuals in the populations. They are "mental constructs that may be used to summarize observed characteristics among relatively homogeneous groups of individuals" (Skinner, p. 71). Empirical typologies stem from a consideration of frequency of individuals according to cross-classifications of variables. The focus is usually on modal types.

Although the central functions of ideal types and empirical types are considered to be distinctive (Wood, 1969), in practice, their use and development often coexist. Both are necessary:

> A key challenge in scientific explanation is the development of constructs that offer both systematic and empirical import. Systematic import refers to the cogency of relationships that connect constructs in the theory plane. On the other hand, empirical import denotes the quality of operational definitions that link theoretical constructs and observable data. (Skinner, 1981, p. 72)

Both empirical and theoretical components are evident in the development of each of Thomas and others' (1968) temperament types: the difficult child, the easy child, and the slow to warm up child. Theory influenced the clinicians' diagnoses which created the clinical sample; data-analytic strategies were applied to clinically isolated groups. As often happens, these three constellations, initially defined by statistically identified dimensions, have

been abstracted to the extent that they now approach ideal types.

The particular course that led to the identification of the three constellations and the resulting temperament classification runs parallel to much of the development of the traditional psychiatric classifications, framed by the medical model. (This observation is quite consonant with the professional backgrounds of the principal investigators of the NYLS, Thomas et al., 1963: four have medical degrees, three of whom have specialties in psychiatry.) The traditional psychiatric classifications have been formulated on a rational basis, usually stemming from thoughtful consideration of a few "classic" cases (Achenbach, 1981; Achenbach & Edelbrock, 1978; Garside & Roth, 1978; Skinner, 1981). They have been based primarily upon symptoms because causes were unknown. The identification of a symptom cluster is called a syndrome. People whose presenting symptoms correspond with a certain symptom profile are considered to be suffering from that syndrome. Psychiatric classifications usually involve either a) the classification of symptoms (or other features) into groups, called syndromes, or less frequently, b) the classification of patients into diagnostic groups, where a category of patients includes those whose symptoms (and other features) correspond with a particular syndrome (Garside & Roth). It is critical to recognize that the identification of a

syndrome does not imply that there is a corresponding group of people distinct from other people, or from people in general (Achenbach; Garside & Roth; Skinner). A classification of behaviors is not necessarily interchangeable with a classification of people.

The end product of the process followed by Thomas et al. (1968) to identify temperament constellations seems more closely related to a classification of behaviors than persons, although not exclusively so. This distinction, that between the classification of behaviors versus people, has not been made in much of the literature on children and temperament (discussed earlier in Chapter I). Statements have been made about types when the data actually pertain to interrelated behaviors. That this distinction is more than an academic exercise becomes clear when placed in the context of screening and intervention programs. These programs do not identify behaviors--they identify children. The systems must, then, be a classification of children, and it should be a good system.

In general terms, to be good means a classification should be a) objective, b) stable, and c) predictive (Cormack, 1971). Establishing methods for evaluation psychological classifications and for determining criteria for a system's adequacy is a relatively recent undertaking (Achenbach & Edelbrock, 1978; Blashfield & Draguns, 1976; Skinner, 1981). There is a general consensus, however, that there

should be sufficient evidence of both internal and external validity. Necessary internal properties of a classification include a) reliability--referring to a group of estimates of consistency of a classificatory label; b) coverage--referring to the "applicability" of a classification to the population for which it was intended; c) homogeneity --where a good system would maximize both within-group homogeneity and between-group heterogeneity; and d) robustness across samples (Blashfield & Draguns; Skinner). External validity of a classification involves "its prognostic usefulness, descriptive validity, clinical meaningfulness of the typal constructs, and generalizability to different populations" (Skinner, 1981, pp. 76-77). Although these properties still suffer from such difficulties as imprecise definitions and measures, they still provide a helpful and stimulating framework from which to begin an assessment of a system.

The necessity of developing a strong classification system is underscored by a consideration of the inherent limitations of any classification system. It is unlikely that any single system will be able to address all needs and purposes simultaneously (e.g., research, clinical, communication, theory, etc.). After all, no classification label can communicate all of the relevant information about a person (Achenbach, 1981; Blashfield & Draguns, 1976). Even when well-suited to the primary purpose for its application,

a system can be limiting. The act of classifying anything gives it a name. Once something has a name, it is apt to be perceived as an actual entity. Such reification can obstruct the observation and understanding of actual objects, processes, and relations. Also, by influencing one's conceptualization, a classification system involving children influences and potentially limits perception and understanding of any child. A classification label generates expectations about a person's behavior. After a person is labelled, predictions may be based on that label and not on the person's actual behavior. The well-known study by Rosenhan (1973) "On Being Sane in Insane Places" exemplifies the dilemma of nomenclature.

It has been recommended by Skinner (1981) that before a classification is used in applied settings, the system should be subjected to standards similar to those required of a psychological test, as specified by the American Psychological Association (1974). Even when the system is empirically sound, there needs to be consideration of whether the contributions of a classification outweigh the negative ramifications, especially when low base rates are involved. This can be quite controversial, particularly within the educational system where a label often becomes a part of a child's permanent record, affecting many decisions. This is of acute concern when the label has a negative connotation. A child can be stigmatized. Braun (1976) has found

that negative information about students influenced teachers more than neutral or positive information. Regarding temperament specifically, on the basis of findings of an extensive research program at UCLA, Keogh (1982b) has stated that "teachers' responses to children in the classroom are mediated by their perceptions of the children's temperamental characteristics" (pp. 274-275). Relevant to types of temperament (although it is unclear whether it is the individual's entire pattern of traits rather than particular dimensions that act as the effective ingredients), Keogh (1982b) found that teachers' decisions were affected by children's temperament and that they viewed children with negative temperament patterns as requiring supervision and as being potential problems. In the actual classroom, Pullis (1979) and Pullis and Cadwell (1982) found that children with more positive temperament patterns received higher teachers' estimates of pupils' ability; also, teachers overestimated the intellectual potential of children with positive temperamental patterns while underestimating that of children with negative constellations. Also from the classroom setting, Keogh (1982b) found that teachers reported a temperament scale to be helpful in pointing out noncognitive areas of individual differences. However, in view of the possible negative consequences, it must be questioned whether that constitutes sufficient justification

for its applied use. The ethical concerns are obvious. They add to the methodological and conceptual concerns mentioned earlier, emphasizing the necessity of further work on the classification of temperament.

CHAPTER III
CLUSTER ANALYSIS

Many quantitative approaches to the empirical identification of types have been introduced since the outset of the NYLS. Although cluster analysis was first discussed in the social sciences 50 years ago (Driver & Kroeber, 1932), it did not attract significant interest until Sokal and Sneath published Principles of Numerical Taxonomy in 1963. In the 20 years since, there has been an "explosive" growth of cluster analysis methods to create classification systems (Blashfield & Aldenderfer, 1978b). In light of researchers' and clinicians' increasing dissatisfaction with available classifications (Skinner & Blashfield, 1982), especially those related to children's behavior (Achenbach & Edelbrock, 1978), it is not surprising that there has been such a response to empirically derived methods. It is assumed that empirically derived classifications will facilitate a more reliable and objective classification of individuals (Skinner, 1981), which should lead to increased understanding. The availability of computers has also fueled the growth of cluster analysis methods (Skinner & Blashfield, 1974; Sokal, 1974).

Cluster analysis is a generic term that refers to a loosely connected family of methods which create

classifications. These methods attempt to form relatively homogeneous groups of subjects called clusters. Although they have various uses, they may be used as descriptive techniques in order to explore the structure of multivariate data sets. Cluster analysis is used in a wide variety of disciplines; it is discussed here as it pertains to psychology. A brief overview of cluster analysis is presented to facilitate the reading of this paper. Several sources (Anderberg, 1973; Blashfield, 1980a, 1980b; Blashfield & Aldenderfer, 1978b; Everitt, 1974, 1979; Skinner & Blashfield, 1982; Sokal & Sneath, 1963) are highly recommended for further detail.

Cluster analysis can be quite useful in classification research. It offers several advantages: the methods are objective and empirical; they can be applied to large data sets which might otherwise overwhelm a person; and they can help "uncover" the multivariate structure of the data. Basically, clustering methods create a classification system by allocating "similar" individuals into the same category, called a cluster. This usually involves four stages (Skinner & Blashfield, 1982). First, data are collected on a large sample. Second, the degree of similarity (or dissimilarity) among every pair of subjects is computed using one of various similarity coefficients. Then, a computer algorithm involving objective criteria is utilized to search for relatively homogeneous subgroups. There are

a variety of different methods which incorporate varying definitions of clusters (Everitt, 1974), similarity (Tversky, 1977), and homogeneity (McQuitty, 1967). (These will be discussed in more detail below.) Lastly, the empirically derived clusters should be validated, both internally and externally (Dubes & Jain, 1979; Everitt, 1974; McIntyre & Blashfield, 1980; Skinner, 1981; Skinner & Blashfield, 1982).

There are numerous cluster analysis methods in the literature. Most of these belong to one of two families: hierarchical agglomerative methods and iterative partitioning methods (Blashfield & Aldenderfer, 1978a, 1978b). Other families of cluster analysis methods include hierarchic divisive methods, density search, factor analysis variants, clumping, and graphics. Many of these techniques have been developed with the focus on certain properties over others. The properties of clusters include their shapes, their dispersion, their location in space, and the size of gaps between clusters (Sokal, 1974). Each of these groups of methods will be described briefly.

Hierarchical agglomerative clustering techniques begin with the computation of a similarity (or distance) matrix between all subjects. The two most similar (or closest) individuals form the first cluster. The method continues by fusing individuals or groups of individuals which are most similar (or closest), culminating when all subjects are

placed in one group. One end-product is a dendrogram or a tree structure which shows the successive fusions of individuals. Single linkage, complete linkage, average linkage, centroid method, median method, and minimum variance method (also frequently called Ward's method) are some of the hierarchical methods of cluster analysis. (There are many equivalent terms for each of these methods; a detailed listing is available in Blashfield and Aldenderfer's, 1978b article.)

These different methods are based upon different definitions of similarity or distance between an individual and a group of several individuals, or between two groups of individuals (Everitt, 1974). They optimize different notions of clusters. All use a matrix of similarity (or distance) between every pair of individuals. Different coefficients of similarity and distance are available. Correlation and distance are popular indices in psychological research, each emphasizing a separate aspect of profile similarity. Pearson product-moment correlation is only sensitive to the shape of profiles; squared Euclidean distance is usually considered when elevation across variables is more crucial than pattern similarity, although it also contains some information about relative profile shape and scatter (Cronbach & Gleser, 1953; Fleiss & Zubin, 1969; Skinner, 1978). The choice of any one measure necessitates a trade-off concerning the type of information utilized.

Although there are several reviews of similarity/dissimilarity coefficients (e.g., Carroll & Field, 1974; Cormack, 1971; Cronbach & Gleser, 1953; Tversky, 1977), there are no established rules governing the selection of coefficients. Certain methods can only use certain coefficients.

The three linkage methods, mentioned above, can be used with any similarity coefficient. In single linkage, the similarity between two clusters is defined as the highest similarity coefficient (or smallest distance) between two individuals, one from each cluster (Wishart, 1978). Initially, there are as many groups as there are individuals. Groups are fused according to the similarity/distance between their most similar/nearest members (Everitt, 1974). Single linkage will produce "straggling" clusters, often failing to partition large samples due to chaining. An example of this can be seen in Figure 3-1. The individuals to the right of the marker on the x-axis have been "chained" on, linked on one by one by their proximity to the last addition. This produces a straggling chain, not a cohesive, compact cluster. Complete linkage is the opposite of single linkage. Here, the similarity between two clusters is the smallest single similarity coefficient between two individuals, one from each cluster, or, if using distance, the distance between clusters is defined as the distance between their farthest pair of individuals. Complete linkage methods tend to find spherical clusters, but the results

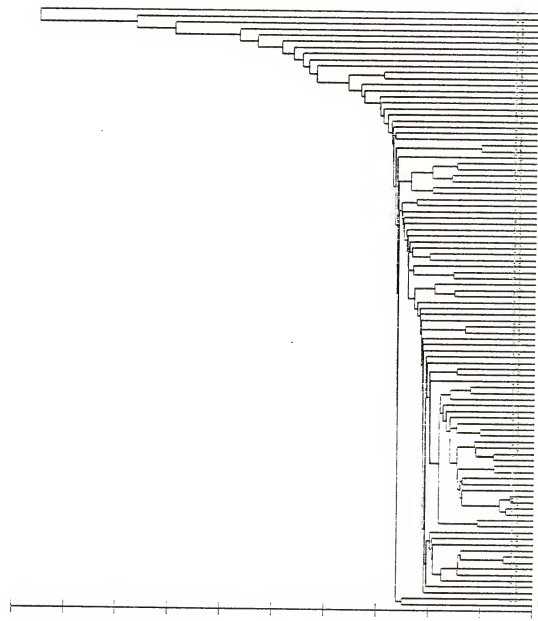

Figure 3-1. Single linkage with squared Euclidean distance, Sample A (n=99).

can be rather irregular because fusions are determined by information for only two individuals, not any measure of the group's structure. Average linkage compares groups by averaging the similarity coefficients or distances between all pairs of individuals in the different groups. "Average linkage tends to find spherical clusters, and is reasonably well behaved" (Wishart, p. 33).

Centroid methods are not meaningful when used with correlation coefficients. Clusters are represented by the coordinates of their centroids. The distance between two clusters is defined as the distance between their centroids. The method fuses those clusters with the smallest distance between their centroids first. This method can also produce chaining, although to a lesser extent than single linkage.

The median method and minimum variance method are only valid when used with distance coefficients. Because the centroid method has the disadvantage of being affected detrimentally if the sizes of two groups to be fused are very different, the median method was developed to be independent of group size (Everitt, 1974). "The distance S(R,P+Q) between any cluster R and the cluster which results from the fusion of P and Q is defined as the distance from the centroid of R to the midpoint of the line joining the centroids P and Q" (Wishart, 1978, p. 33). Median can also chain for large populations.

The minimum variance method involves the error sum of squares which is defined as the sum of the distances from each individual to the centroid of its parent cluster (Wishart, 1978). The method combines the two clusters whose fusion produces the least increase in the error sum of squares. Minimum-variance spherical clusters are found.

The major disadvantages to this family of techniques are 1) there is no step for rectifying a "bad" or ineffectual fusion which will affect all subsequent fusions, 2) the tendency for some of these methods to fail to initiate clusters because of chaining, and 3) because the procedures start with as many clusters as subjects and proceed to fuse clusters until all are in one, the researcher must decide, without the help of established rules (Everitt, 1979), the most appropriate number of clusters.

Hierarchical divisive techniques are similar to hierarchical agglomerative methods, although they work in the reverse order. The whole set of individuals is first divided into two groups. Then, each subset can be divided into further subsets, until there are as many subsets as subjects. These techniques are used primarily in biology, ecology, and anthropology (Blashfield & Aldenderfer, 1978b) and suffer the same disadvantages as do the hierarchical agglomerative techniques.

Solutions produced for the hierarchical methods are often presented graphically. Graphic output provides a

two-dimensional representation of the fusions of individuals. For example, the minimum spanning tree utilizes a branching tree to represent the structure of similarities among subjects. Although there has been little formal study of their utility, many investigators find them useful for "exploring" their data, identifying outliers, and deciding how many clusters there are. Also, an investigator can use a plot of subjects with the two principal components defining the axes and with outlines surrounding the members of each of the clusters of a solution (Blashfield, personal communication). Other graphic theoretic methods have been proposed (e.g., Everitt, 1979), although they do not appear very much in the literature (Blashfield & Aldenderfer, 1978b).

The iterative partitioning methods, unlike the hierarchical techniques, allow relocation of any misassigned subjects to a more appropriate cluster. These methods begin with a predetermined classification, or partition, then employ some iterative process to revise the classification. They "alter cluster membership so as to obtain a better partition. . . . The various algorithms which have been proposed differ as to what constitutes a 'better partition' and what methods may be used for achieving improvements" (Anderberg, 1973, p. 156).

Partitioning methods start with an initial classification of the sample and then apply an iterative relocation

procedure to reassign individuals until there is no better solution according to some specified optimum. If the same solution is obtained from several different starting classifications, the probability is greater that the global solution has been achieved (Wishart, 1978). The investigator decides upon the appropriate number of groups, called k, in the data set. For many applications an investigator knows how many groups are of interest. Most of these methods find solutions for a fixed number of clusters, although there are few that allow for variability (Anderberg, 1973). The initial classification of the sample provides the first cluster centroids to which the subjects are assigned. The clusters' centroids are adjusted as their membership is altered.

A popular iterative partitioning method is the k-means partitioning method, which denotes the process of assigning each person to that cluster (of k clusters) with the nearest centroid (mean) (Anderberg, 1973). Usually proximity is determined by squared Euclidean distance. Once every subject has been assigned to a cluster, all clusters are checked for any members who should be reassigned (i.e., they are closer to the centroid of some other cluster). This decision usually depends upon the optimization of some selected clustering criterion statistic, usually error sums of squares. This process is repeated iteratively until a stable solution is arrived at.

Besides being able to reassign subjects, these partitioning techniques have another advantage over hierarchical techniques; they do not require the calculation and storage of a similarity matrix, so they are able to handle much larger data sets (Anderberg, 1973). However, there are disadvantages. The major ones are that the initial partition often affects the cluster solution and that an exhaustive search of all possible partitions is enormously expensive. Also, little is known about the effects of initial partitions, the type of pass for assigning individuals to particular clusters, and the various optimization criteria (Blashfield & Aldenderfer, 1978b).

The methods described above are the most frequent ones found in psychological research. Other techniques are available, although they have rarely been used in psychology with applied data. Therefore, their utility and characteristics are not yet well understood. (Although they are included in this overview of cluster analysis methods, they are not of much importance to this particular study, except for the sake of curiosity, in terms of how they partition the data.)

For instance, there are density search techniques, such as Mode (Skinner & Lei, 1980; Wishart, 1969). If individuals are depicted as points in hyperspace, there should be regions that are very dense, separated by regions of relatively low density. These techniques are aimed at finding these dense areas.

Variants of factor analysis, especially Q-type analysis (Cattell, 1952), have been applied in psychological research. According to this method, correlations between individuals are calculated, instead of the correlations between variables that are characteristic of "normal" factor analysis. Then, the usual methods of factor analysis are applied to the matrix of correlations. Individuals are assigned to clusters according to their factor loadings on the extracted "factors." This method has stirred considerable controversy. The major criticisms involve the underlying assumptions of these methods. Fleiss and Zubin (1969) and Sawrey, Keller, and Conger (1960) have questioned the meaning of correlations, asking what it means, in terms of similarity, to say that two individuals are highly correlated. More specific to factor analysis, there has been criticism of the constraints of linearity (Everitt, 1969; Fleiss & Zubin).

> It is probably universally agreed that a factor analysis is an idle exercise when performed on data for which there is little chance that an underlying linear model is tenable. If so, then what does one make of Q-factor analysis? No adequate exposition of the applicability of the linear model to people and types appears to exist. (Fleiss & Zubin, p. 238)

On more pragmatic grounds, factor analysis has been criticized because of its poor performance in practice (Blashfield, 1977).

Clumping techniques allow overlapping clusters, unlike those described above, where cluster solutions are usually disjoint. The primary application for these techniques has been in the area of language studies, where words must be able to belong to several groups because they have several meanings (Everitt, 1974). Given their limited usage, these techniques' characteristics are not yet well understood (Jardine & Sibson, 1968).

Although the basic notions behind clustering methods are rather simple, the utilization and interpretation of solutions are quite complicated. The cluster analysis literature, not having a firm theoretical base, suffers from confusion at methodological and conceptual levels. Even the terminology reflects and perpetuates confusion. The inconsistent use of equivalent terms (for example, elementary linkage, nearest-neighbor, space-contracting, and connectedness methods are all equivalent terms for single linkage methods) and the fragmentation of terminology into jargon hamper communication (Blashfield, 1980a; Blashfield & Aldenderfer, 1978b). Communication difficulties impede improvement, comprehension, and correct usage.

Other issues are even more problematic for the user of cluster analysis:

> The novice user of cluster analysis soon finds that even though the intuitive idea of clustering is clear enough, the details of actually carrying out such an analysis entail a host of problems. The foremost

> difficulty is that cluster analysis is not a term for a single integrated technique with well-defined rules of utilization; rather it is an umbrella term for a loose collection of heuristic procedures and diverse elements of applied statistics. The actual search for clusters in real data involves a series of intuitive decisions as to which elements of the cluster repertory should be utilized. (Anderberg, 1973, p. 10)

Without a standardized procedure for a "correct" cluster analysis, the user must make decisions about each of the following elements of a cluster analysis (Anderberg): choice of data units, choice of variables, what to cluster, homogenizing variables, similarity measures, clustering criterion, algorithms and computer implementation, number of clusters, and interpretation of results. Each element may affect the cluster solution. Choice of algorithms and interpretation of results are discussed below; the other elements are left to the next chapter for discussion.

Although cluster analysis methods are touted as being able to "find" clusters, they are, in fact, techniques that impose structure, that fit the data to the technique. A cluster analysis can find structure where none exists, even for random data (Dubes & Jain, 1979). Different methods applied to the same data can produce different solutions (Bartko, Strauss, & Carpenter, 1971; Everitt, 1974), as can the choice of similarity measure (Edelbrock, 1979). To complicate matters even further, different software for the same method can produce different results (Blashfield, 1977).

The combination of these problems culminates in a question: Which of the innumerable techniques are the good ones? In most applied research, cluster analysis methods are being used on data sets for which the "true" structure is not known. It is impossible in this situation to know if a particular method generates an accurate solution. A comparison of the solutions of different methods may reveal similar results across methods, i.e., generality. However, generality is no guarantee of accuracy, much in the same way that reliability is no guarantee of validity.

This dilemma has led to an approach known as Monte Carlo research. In this work, artificial data sets are generated to have a particular structure. Then, the solutions obtained by a certain technique may be compared with the "true," i.e., generated, structure. Everitt (1979) summarized this work as follows:

> In general the results . . . indicate that 1) no single method is best in every situation, 2) the mathematically respectable single linkage is, in most cases, the least successful for the data used, and 3) group average clustering and a method due to Ward (Ward, 1963), do fairly well overall. (p. 173)

(Equivalent terms for those are average linkage and minimum variance, respectively.) These methods, especially average linkage used with Pearson product moment and minimum variance used with Euclidean distance, have received further

support from the studies of Milligan and his colleagues (Milligan, 1981a, 1981b; Milligan, Soon, & Sokol, Note 2). Neither one emerges as always superior to the other.

A variety of factors have been shown to affect cluster recovery. Some methods are superior under certain conditions. In a review article Milligan (1981b) describes three factors which seem to determine whether average linkage or minimum variance gives better recovery. The first factor concerns the choice of similarity measure used to form the initial matrix. The minimum variance method was superior when Euclidean distance was used; the average linkage method was equivalent to the minimum variance method when the Pearson correlation coefficient was used. The second factor concerns the treatment of outliers or entities between clusters. Although there are some inconsistent findings, it seems that when there is a requirement of total coverage, i.e., that all individuals of the sample must be assigned to one of the solution groups, that minimum variance may give better recovery. The third factor involves the amount of cluster overlap. When clusters overlap, the minimum variance method gave the best recovery, especially as the degree of overlap increases. More research is necessary before these problems are resolved. In the meantime, it is suggested (e.g., Everitt, 1974, 1979) that several methods be used to cluster a data set.

> At present, the best approach is to use the more reliable clustering procedures to produce several sets of clusters and then to compare the sets to determine which individuals always cluster together. (Maurer, Cadoret, & Cain, 1980, p. 523)

In this manner, a user has a solution. Yet, given the fact that cluster methods can "find" solutions where no structure exists, it is necessary to evaluate the accuracy of the solution, i.e., its validity. McIntyre and Blashfield (1980) discuss two characteristics of a good cluster solution: 1) it is stable (replicable) across multiple data sets, and 2) it matches the "true" structure--it is accurate. Because the "true" structure is what is unknown and of interest in applied research, McIntyre and Blashfield developed a procedure which estimates a solution's accuracy by measuring its stability. It is called the nearest-centroid evaluation technique.

The basic steps in the nearest-centroid procedure are as follows (McIntyre & Blashfield, 1980):

1. Two independent samples of multivariate data (Sample A and Sample B) are randomly selected.
2. Sample A is cluster-analyzed.
3. The centroid vectors for each cluster are calculated.
4. Sample B is cluster-analyzed.
5. The squared Euclidean distance for each of Sample B's objects from each of the centroids of Sample A is calculated.
6. Each object in Sample B is assigned to the closest centroid vector.

> 7. The agreement between the nearest-centroid assignment of the previous step and the cluster results of step 4 is measured with the kappa statistic. This is called "agreement kappa" . . . and is an index of the goodness of the cluster solution. (p. 228)

Accordingly, a variety of cluster analysis techniques in combination with various similarity measures can be applied to Sample A. A "best" solution can be selected from these solutions. If the "best" solution for Sample A has generated valid types, one would expect that a replication of the clustering method chosen as best would generate similar types for Sample B. They found in Monte Carlo studies that the degree of agreement between the nearest-centroid assignments and the results of the cluster analysis of the second sample has merit as an estimate of the solution's stability.

McIntyre and Blashfield (1980) measured the degree of agreement between the two classifications of the second sample with a statistic called agreement kappa. Kappa (Cohen, 1960; Fleiss, Cohen, & Everitt, 1969; Hubert, 1977) is a statistic which can be used to measure the agreement between two classifications of the same data set. Kappa ranges from 0 (no agreement beyond chance level) to 1 (perfect agreement). Agreement kappa provides not only a direct estimate of stability (beyond chance levels) but also "an indirect estimate of how well the minimum variance cluster solutions matched the actual cluster structure of

the data" (McIntyre & Blashfield, p. 236). This technique serves as a procedure for internal validation. Once there is some evidence of a cluster solution's validity, interpretation of the valid solution's clusters, i.e., the empirical types, may proceed.

Despite the associated problems, cluster analysis remains a potentially useful technique for classification research due to its relative objectivity and its empirical derivation. However, it is important to remember that cluster analysis is primarily a tool for discovery. It is designed for heuristic aims, not hypothesis testing.

CHAPTER IV
CLUSTER ANALYSIS AND TEMPERAMENT

In Chapter I temperament was introduced as delineated by Thomas and Chess and their colleagues (1963, 1968, 1977, 1980). The discussion concentrated on their three temperament constellations, namely, the difficult child, the easy child, and the slow to warm up child. A review of a) the process by which each of these constellations or types was identified and b) the research concerning the empirical basis for these types (most of which actually pertained to dimensions, not types) raised serious concerns about their methodological adequacy for application. It was concluded that more research on temperament type is indicated, prior to its application.

In Chapter II the discussion focused on types within the framework of classification theory and systems. It was suggested that the Thomas and Chess constellations, as currently conceptualized, can be considered ideal types. Identifying empirical types can, then, serve as a means for assessing the validity of ideal types (Skinner, 1981; Wood, 1969). Quantitative techniques for the empirical identification of types have proliferated since the outset of the NYLS. These quasi-statistical methods, called cluster analysis techniques, were introduced in Chapter III.

It is the purpose of this study to employ cluster analysis techniques to "look at" temperament data for natural groupings, i.e., types, of children. Cluster analysis permits a relatively objective method for identifying types. These empirical types, in comparison to the three identified by Thomas et al. (1968), are used to address the issue of temperament type.

Using cluster analysis to study temperament is not without precedent. There have been two published studies (to the author's knowledge) that have employed clustering techniques to address the question of whether there are distinguishable groups of children on the basis of temperament.

McDevitt and Carey (1978) developed a parent rating instrument to measure temperament in three- to seven-year old children, called the Behavioral Style Questionnaire (BSQ). They reported having performed a "person cluster analysis" on the standardization sample using a statistical program by Tryon and Bailey (1970). They selected 10 individuals whose profile patterns fit the theoretical definitions (Thomas et al., 1968) of difficult, easy, and slow to warm up children (10/group). These 30 formed the working definitions of the clusters, to which the computer program assigned as many as possible of the remainder of the sample. Although their findings of similar prevalence of types and lack of sex differences were presented as evidence of the

BSQ measuring the same temperament characteristics as described by Thomas et al., the methodology is inadequate to support the question of types. The computer algorithm provided an identifying technique, more than a clustering technique (Everitt, 1974). They assigned individuals to predefined groups; they did not "discover" groups as existed naturally. Also, they employed only one clustering method. Inasmuch as any method will impose structure and yield a solution, additional methods are necessary to provide minimal evidence of the stability of a solution. Also, the authors did not publish sufficient information to enable replication of their work (e.g., their criteria for determining stopping points for assigning subjects to the three groups). They used only six of the nine variables in this assignment. Although the authors do not discuss their rationale, it seems this decision stemmed from considering only the dimensions used to define these groups. Whether the use of all nine might have greater discriminatory information remains of interest.

The second study was conducted by Maurer, Cadoret, and Cain (1980). They employed several clustering techniques to address the issue of temperament types. Although commendable for its attention to the stability and validity of the clusters, methodological problems limit the meaning of their findings. A major problem is their use of an instrument that had not been standardized. Another problem is

the large number of variables (61) relative to number of subjects (N=162). Although there is no hard rule, the rule of thumb suggests that as a minimum, the number of subjects should be 10 times the number of variables. Also, boys and girls were clustered separately, making it impossible to separate out whether differences in cluster solutions across samples reflected a lack of stability of the clusters or sex differences or both. Also, the authors did not report sufficient information concerning the methods used, the choice of similarity measures, and the decisions pertaining to the number of clusters which impedes the reader from evaluating the clustering procedure or replicating it (Blashfield, 1980b). The study is also limited by its retrospective design (Yarrow, 1963). In addition to the methodological problems of these two studies, this study also attends to the practical problems of cluster analysis. It is the purpose of this chapter to discuss these practical aspects (e.g., Anderberg, 1973; Blashfield, 1980b; Dubes & Jain, 1979; Everitt, 1974, 1979) as they pertain to this study.

Choice of variables. One practical consideration is the choice of variables. Thomas et al. (1968) identified the three temperament constellations by comparing children's scores on nine different dimensions of temperament. The scores were derived from interview data obtained from the children's parents. This particular method has the

disadvantages of being time-consuming, costly, and relatively difficult to replicate because of subjective influences in the interviewing and in the scoring of interview data. Work on the measurement of infant temperament proceeded quite swiftly (Brazelton, 1969, 1973, 1983; Carey, 1970, 1972b; Carey & McDevitt, 1978a, 1978b, 1980; Egeland & Deinard, 1978; McInerny & Chamberlin, 1978). Less has been done with young children. There were three major rating techniques for assessing the temperament of young children that were available at the time this study was begun. These were Thomas and Chess' (1977) Temperament Questionnaire, Rowe and Plomin's (1977) Colorado Children's Temperament Inventory, and McDevitt and Carey's (1978) Behavioral Style Questionnaire (BSQ). Given the relative strengths of the BSQ in terms of its empirical basis (moderate-to-high reliabilities, preliminary evidence for construct and predictive validity, a large N, relatively long retest period, and a high return rate of the questionnaires), the BSQ was selected to be the rating instrument for this study. (For more details concerning the development of this instrument, the reader is referred to the articles by Carey, Fox, and McDevitt, 1977, and McDevitt and Carey, 1978.) Each item is relevant to one of nine temperament categories, developed to measure the temperament dimensions as delineated by Thomas et al. (1963). Scoring procedures yield an averaged

score for each of the dimensions. These nine scores are the variables that are clustered.

As mentioned in Chapter I, using parents' ratings to study children's temperaments has been questioned as a valid procedure. The most recent opinions in this controversy tend to favor parent ratings as moderately reliable and valid when the instrument items request behavioral descriptions, usually in terms of relative frequency, for the child's recent behavior. Items that call for retrospection and/or interpretation of behavior tend to lower an instrument's reliability. The BSQ calls for frequency ratings of the child's recent behavior.

Given the advantages of parent rating instruments over semi-structured interviews, and given the BSQ's sufficient empirical basis as well as its theoretical structure being that of Thomas et al. (1963, 1968), it was selected as the most appropriate measure. However, it is agreed that, technically, what is being measured is children's perceived temperament (Keogh, Kornblau, & Ballard-Campbell, Note 3).

Selection of subjects. Selection of subjects is another element of any cluster analysis. One impetus for this study was Thomas and Chess's promotion of the temperament construct for use in screening and intervention programs without there having been established, in the

author's opinion, a sufficient empirical basis to justify its applied use. The aim of many screening and identification programs is preventative. Children who are at-risk for developing problems are identified and treated in a manner designed to decrease that risk. Thomas and Chess (1977) and Keogh (1982b) have presented the difficult and the slow to warm up characteristics as identifiers of at-risk populations. The particular target group has been young school-age children, at-risk supposedly for school-related problems. Therefore, this study has selected the same age group as that typically included in such a screening program: the subjects are three- to seven-years old.

After the BSQ was selected as the instrument of choice, it was decided to select subjects so to replicate as closely as possible the composition of the BSQ standardization population, in terms of age, sex, and SES level. Cluster analyses do not require random and independent selection to be valid, although those conditions facilitate generalizing from a study's sample to the population of interest (Anderberg, 1973).

What to cluster. What gets clustered can also affect cluster solutions. In this study, individuals are clustered on the basis of their temperament scores. Neither Thomas et al. (1968) nor McDevitt and Carey (1978) utilize all nine temperament dimensions to identify children as difficult, easy, or slow to warm up children. It remains a

question whether the "extra" dimensions would be useful in discriminating between types of children. Cluster analyses are performed on both the six dimensions used by McDevitt and Carey to assign children to "diagnostic clusters" and the full nine dimensions. The two sets of solutions will be compared for their utility in initiating meaningful clusters.

Standardization. Another debated issue is that of standardization (Everitt, 1974, 1979), which Anderberg (1973) calls homogenizing the variables. This is a critical issue when variables are not measured in equivalent units (for example, favorite color and age in years). However, the nine temperament variables of the BSQ are measured in similar units. Judging from the standardization sample (McDevitt & Carey, 1978), it is expected that standardization will not be considered to be necessary once this sample's means and variances are inspected. Not standardizing can be preferable because standardization can dilute the differences between groups, by reducing the contribution of the variables with large variances to those with smaller variances (Anderberg, 1973; Everitt, 1974, 1979). This, obviously, counteracts the aim of clustering. Also, as Everitt (1974) points out, some clustering techniques give different solutions when data have been standardized.

Clustering. In order to provide information for the evaluation of a clustering solution, the nearest-centroid technique (McIntyre & Blashfield, 1980), as discussed in Chapter III, is followed. Accordingly, a large sample will be randomly divided into two groups, Sample A and Sample B. Sample A is the derivation sample, according to this technique's cross-validation paradigm. A variety of clustering techniques in combination with various similarity coefficients are used to cluster Sample A. Given the sensitivity of clustering methods to extremes and variability, attention is paid to "outliers." A "best" solution is selected from these solutions by considering a solution's replicability across techniques, consistency of membership, and "clinical meaningfulness." Wishart's (1978) CLUSTAN computer program package has been selected because of its superior flexibility and versatility (Aldenderfer & Blashfield, 1978).

Number of clusters. Another practical problem is that of deciding the most appropriate number of clusters present in the data. There is no clear indicator for the number of clusters (Anderberg, 1973; Cormack, 1971; Everitt, 1974, 1979). (This is not too surprising, considering the difficulty in specifying the definition of a cluster.) For hierarchical techniques, an examination of the dendrogram for large changes between fusions is useful (Everitt, 1974). The clustering coefficients which relate to the amount of

variance or similarity accounted for at each step of the clustering process can also be examined to assist in determining the number of groups. When these coefficients are graphed, it is often possible to see "jumps" in the values which are out of proportion to previous changes. Looking at the members of the cluster before the jump, in other words--when the cluster appears more homogeneous (smaller within-cluster variance)--provides more information about the appropriate number of clusters. The means for the clusters can be compared across solutions, also. As is apparent, there is a considerable degree of subjectivity involved, which underscores the necessity for validation.

Validation. The nearest-centroid evaluation technique (McIntyre & Blashfield, 1980) provides an internal validation procedure. Internal validation refers to the evaluation of a clustering solution by itself, without concern for the subject matter (Dubes & Jain, 1979). It belongs to a group of internal validation procedures called data manipulation procedures. These are techniques designed to assess the generality of a clustering solution. Although there are several possible manipulations, this one involves the split-sample replication. It provides a method for assessing the validity of the solution, even if minimal in terms of evidence of validity. In this nearest-centroid evaluation technique, the agreement between the two classifications of Sample B is measured. Significant agreement

reflects stability, which provides an estimate of the solution's validity. If a solution is found to be valid, it can then be interpreted.

Given the abundant confusion in the literature about cluster analysis, it is important to communicate any clustering work as specifically as possible. Some of the "intuitive" judgments are difficult to record, yet much of the procedure is more tangible. To this aim, the propositions regarding the use of cluster analysis suggested by Blashfield (1980a) are followed.

CHAPTER V
METHOD

Subjects

Questionnaires were distributed in child care centers (described below) to the parents of three- to seven-year-old children. Participation was voluntary. Informed consent was obtained in accordance with the American Psychological Association's guidelines. No incentives were provided. Two hundred fifteen questionnaires were returned: 208 had been completed, and 7 were blank. Of the 208, 8 were removed from the sample due to an excessive number of items left blank, which was defined as 80% or less of a category's items completed. Of the remaining 200, there were 92 for girls and 108 for boys. The age distribution of the sample was as follows: 3-0 to 3-11 n=56 (28 girls, 28 boys), 4-0 to 4-11 n=51 (21, 31), 5-0 to 5-11 n=53 (28, 25, 6-0 to 6-11 n=29 (13, 16), and 7-0 to 7-11 n=11 (3, 8).

Of the 200 parents who completed the questionnaires, 162 were mothers of a child between 3 and 7 years old, 35 were fathers, and 3 were another relation (grandparent or step-parent) who was the child's guardian.

Most of the raters (70.5%) were married; 25% were separated or divorced; 2.5% were single parents; and 2%

were widowed. The raters' ages ranged from 21 to 58 years, with most in their 30's (67.5%). In 42 instances, the spouse's age was left blank. (For most of these, the rater was divorced or a single parent.) When provided, the spouse's age ranged from 21 to 53 years, with the majority in their 30's (59%).

Most of the raters were Caucasian (91%). Of the remainder, five were Afro American, six were Asian, two were Hispanic, three were "Other," and two left the item blank. Spouses were described in similar proportions.

All social classes were represented in the sample with a preponderance considered middle class. In terms of the highest academic degree earned, 1 rater had not completed high school, 78 had high school degrees, 17 had associate degrees, 50 had bachelor degrees, 44 had advanced degrees, and 10 checked "other." Similarly, 4 spouses had not completed high school, 67 had high school diplomas, 10 had associate degrees, 32 had bachelor degrees, 42 had advanced degrees, 2 had "other," and, in 43 cases, the information was left blank.

Occupationally, most of the raters worked outside the home; 6 were full-time students; 3 were self-employed; 11 were manual laborers; 54 were in some sort of skilled labor (e.g., word processor, foreman, nursing assistant); 77 were in "white collar" jobs that required college-level training; and 21 were professionals (e.g., lawyer, doctor, executive).

Information about the spouse's occupation was not provided on 50 questionnaires. Of the rest, 2 were students; 4 were self-employed; 25 were manual laborers; 26 were skilled laborers; 52 were in "white collar" jobs; and 31 were professionals.

In terms of annual family income, 3 people did not provide the information. Of the rest, 27 earned up to $10,000, 26 between $10,000-15,000, 37 between $15,000-20,000, 53 between $20,000-30,000, 42 between $30,000-50,000, and 12 earned more than $50,000.

In terms of family size, most families had 1 (41%) or 2 (43%) children; 26 families (13%) had 3 children; 3 families had 4; and 2 families had 6. The birth position of the described child was as follows: 83 were only children; 47 were oldest; 7 were middle of 3 or more; and 63 were youngest.

Selection of child care centers. A list of child care centers (e.g., nursery schools, daycare centers, after school care, etc.) for the western and southern suburbs of Boston, Massachusetts, was obtained from the State of Massachusetts' Office for Children. A similar list was obtained from the State of New Hampshire for Portsmouth and the rest of the seacoast region. This area lies 50 miles north of Boston, Ma., and is quite diversified in its population, including rural, university, and urban segments.

Initially, a letter of general introduction was mailed to the director of each of 27 centers. These initial centers were chosen a) on the basis of being in session during that time of year (summer and early fall), and b) as composing a representative sample of the existing variety of centers (in terms of number and age of enrolled children, cost, and type of program). As data collection continued, six other centers were contacted, such that data were collected during all seasons (so not to introduce a bias stemming from the inclusion of only those centers will full-year or summer-only programs).

As a follow-up to the introductory letter, each director was telephoned to discuss the study and to determine whether there was any interest in participating. Of the 33 places which received the initial letter, it was possible to contact 27 directors: 16 agreed to meetings; 6 said "maybe later if you're stuck," and 5 declined outright. Questionnaires were distributed at centers in order of agreement. Thirteen centers were necessary to obtain the desired number of completed questionnaires.

Each of the directors who declined immediately offered an explanation. These can be grouped as a) research-related concerns, such as already being involved in a project or having had an unpleasant previous experience, b) administrative concerns, such as being overextended or cost-accountable, and c) people-related concerns, such as having

very few children between the ages of three and seven years, few English speaking/reading parents, and low reliability of parents in terms of returning forms.

If the director expressed an interest in participating, he/she was sent a set of the materials that each parent would receive, and an appointment to meet was made. At the meeting the general purpose and procedure were explained, and any questions were answered. All directors who consented to this meeting went on to give permission for their center to participate.

The general procedure was explained as follows: 1) the director would provide a list of names of the parents of three- to seven-year-old children; 2) a set of materials would be distributed at the center to each child's parents; and 3) the director would remind parents to return the questionnaires and would inform the investigator when there were materials to be collected. There were slight procedural variations within centers, depending upon the usual method of distributing information to parents as well as the degree of familiarity between the director and the parents. Sometimes the director preferred to address the envelopes herself (all participating directors were female) in order to safeguard the parents' confidentiality. Some directors distributed the questionnaires personally; some allocated the responsibility to the teachers. The variations between centers are thought

to be insignificant, although a systematic bias cannot be ruled out.

Materials

Parents were given a nine-page anonymous questionnaire.[2] The first page was a cover letter that briefly explained the purpose of the study, requested the one-time participation of the parent, and identified the researcher. It was emphasized that participation was voluntary and unrelated to school enrollment. The second page was a personal data sheet which requested demographic information. The next seven pages were comprised of the Behavioral Style Questionnaire (BSQ) by McDevitt and Carey (1978). The first of those pages asked for the child's age and sex, the rater's relationship to the child, and it gave instructional

---

[2]Initially, questionnaires were not designed to be anonymous, with a thought toward the possibility of a follow-up study. However, directors were pessimistic about parental participation in something that could potentially allow their child to be identified in an undesirable way. Several directors expressed their own reluctance to participate unless materials were anonymous. In fact, when tried at four centers, return rates were low (7-24%), some demographic information was omitted in almost every questionnaire, and most of the questionnaires that were returned had an unacceptably high number of items left blank, often rendering it impossible to score one or more of the temperament dimensions. It was decided to change the materials to make participation anonymous. Data collected prior to this decision were discarded. Center selection was restarted.

guidelines for completing the 100 questionnaire items, which filled the next six pages.

Briefly, the BSQ is a paper-and-pencil questionnaire, designed to utilize parents' observations to assess nine dimensions of children's temperament, aged three- to seven-years. Raters are instructed to quickly rate every item according to their own observations of the child's recent and current behavior. Raters are presented 100 statements and asked to mark the space that tells how often the child's recent and current behavior has been like the behavior described (space 1 = almost never, 2 = rarely, 3 = usually does not, 4 = usually does, 5 = frequently, 6 = almost always). The estimated average completion time for first completion is 25 minutes.

Items were written for each of the nine dimensions of temperament as delineated by Thomas et al. (1963, 1968). Items include high and low extremes of each of the dimensions. For example, for activity, Item 70--"The child runs to get where he/she wants to go."; Item 26--"The child sits quietly while waiting." Readers interested in further details concerning the development of this instrument are referred to McDevitt and Carey's article (1978). The BSQ can be purchased from Dr. William Carey.

The BSQ is scored using the BSQ scoring sheet which yields an averaged category score for each of the nine temperament dimensions. Each individual's nine scores can

then be entered on a BSQ profile sheet to classify each child according to his/her "diagnostic cluster" or temperament type. There are five possible "diagnostic cluster" assignments. They are described as follows:

1. Easy--typified as rhythmic, with a tendency to approach new situations, adaptable, mild in intensity of responding, and positive in mood;
2. Difficult--typified by arrhythmicity in biological functions, withdrawn, slowly adaptable, intense reactions, and negative mood;
3. Slow to warm up--typified by a low activity level, withdrawn, slowly adaptable, mild intensity, and negative mood;

Intermediates--all others;

4. Intermediate highs--above the mean for many of the categories, but not enough so to qualify as difficult; and,
5. Intermediate low--mostly below the mean but not easy or slow to warm up.

McDevitt and Carey's assignment rules are specified using category means and extent of variation about the mean as criteria.

Procedure

Child care centers were contacted in the manner described above. Within these centers, 263 questionnaires were distributed. Two hundred fifteen were returned, for an overall return rate of 81.7%. Of these, seven were returned blank, indicating as per the directions, that the parent did not wish to participate. The overall rate for participation, therefore, was 79.1%. The remaining 208 questionnaires were coded and tabulated for demographic information. Then each was scored, resulting in nine temperament category scores. From these, a "diagnostic cluster" was determined for each individual.

An inspection of the 208 completed questionnaires was performed to determine whether or not any items were left blank. If more than two items within a category were blank or less than 80% of a category's items were completed, the questionnaire was discarded from the sample. Eight were discarded (range: 3-17 unanswered items, mean = 9.1). The number of times each item was left blank and written comments were considered for potential relevance to the BSQ instrument.

To compare this sample with McDevitt and Carey's BSQ standardization sample, statistical tests were performed to compare the two groups' means and variances. Also, the two samples' distributions across BSQ diagnostic clusters were compared.

For this sample, analysis of variance and of covariance procedures (using SAS) were conducted to test for any main effects of sex and age and sex-age interactions across the categories. Also, a 9 x 9 intercorrelation matrix of the categories was computed to address the issue of dimensionality.

A computer program (using SAS) was used to divide the sample in a random fashion into two groups, Sample A and Sample B. Statistical tests were performed to test for any significant differences between these two groups in terms of sex and age.

All of the cluster analyses were done using CLUSTAN 1C, a package of clustering programs developed by Wishart (1978). The nearest-centroid evaluation technique (McIntyre & Blashfield, 1980) was followed, as explained in Chapter III.

Clustering Sample A

A wide variety of techniques and coefficients was used to "explore" the structure of Sample A. The object was to use many and diverse methods and then select a "best" solution. Inasmuch as the literature supports the minimum variance method with squared Euclidean distance and the average linkage method with correlation as the best methods for recovering the structure of data sets,

primary attention was focused on these methods and their solutions. Other methods were also tried, however.

The first cluster analysis was single linkage with squared Euclidean distance. Although this clustering method is often not very useful because it tends to chain on individuals, rather than cluster them, the chaining itself can be useful for identifying outliers. First, the dendrogram is inspected for the last linkage of a cluster that consisted of more than a single individual. The dendrogram for Sample A was presented in Chapter II (see Figure 3-1). The last instance of two or more individuals being linked can be found immediately to the left of the marker. All of the individuals to the right of the marker have been chained on. These individuals are ones to be considered as outliers. By graphing the linkage coefficients for these individuals, a sudden change in the slope, signaling a change in the usual similarity between clusters, can be used to suggest which of these individuals may be outliers. Then, as additional confirmation that these individuals do, in fact, "lie" outside of the majority, the z scores were computed for the total sample (N=200). If an individual had two or more z scores that were larger than two standard deviations, he/she was confirmed to be an outlier. In keeping with the guiding rule of conservatism for work with clustering techniques, and in view of these techniques' sensitivity to outliers, any confirmed outliers were removed from Sample A, forming

Sample A-R (Revised). To check whether the removal of outliers was necessary or useful, a variety of methods was applied to both Sample A and Sample A-R. Solutions were compared. There were better cluster solutions in terms of distinctiveness, generality across methods, clarity of interpretation or meaningfulness when the outliers had been removed from the sample. Therefore, clustering continued with Sample A-R.

In much the same way, the effects of standardizing the data (versus using raw data) were compared. So were the effects of using six of the temperament dimensions versus all nine. It was felt that better solutions were obtained with the raw data for all nine of the dimensions. All solutions for Sample A-R and Sample B-R (Revised) were generated using the raw data for all nine dimensions consequently.

The following hierarchical agglomerative methods were used once with Pearson product-moment correlation for the similarity measure and once with squared Euclidean distance for the similarity measure: single linkage, average linkage, complete linkage, and McQuitty's similarity analysis. Some hierarchical agglomerative methods are only meaningful when distance coefficients have been used to compute the similarity matrix. Of these, the following were used with squared Euclidean distance: median, centroid, and Ward's minimum variance method. Horizontal dendograms were produced for each of the hierarchical procedures, except

median and centroid. With most clustering methods, it is necessary to specify the minimum and maximum number of clusters which are of interest to the user. Visual inspection of the dendrograms suggested a maximum of five and a minimum of three.

Iterative partitioning methods were also used to cluster analyze the data. The k-means partitioning method was used with error sum of squares as the similarity coefficient. When an individual was being considered for the "goodness" of its assignment to a particular cluster, relative to all others, it was removed from that cluster for the relocation test so not to affect the parent cluster's centroid. The population gets scanned until no objects are relocated during one full scan, which means that an optimal solution has been obtained for the selected parameters, i.e., a local optimum. (If no stable cluster resulted within 10 iterations, scanning was halted to minimize cost.) Then, the two most similar clusters were fused, and the relocation phase was repeated. This was stopped once the sample was reduced to two clusters.

Several different initial classifications of the sample were used. The method started with the children being assigned to 1) 1 of 5 groups, depending upon their BSQ diagnostic cluster, 2) 1 of 10 groups, assigned randomly, 3) 1 of 5 groups, according to their cluster assignment from the 5-cluster solution of the minimum variance technique

with squared Euclidean distance, and 4) 1 of 4 groups, according to their cluster assignment from the 4-cluster solution of the minimum variance technique. These last two initial classifications allow for the relocation of any poorly assigned individuals by the hierarchical method.

In addition to the hierarchical and the iterative partitioning methods, a density search procedure called Mode was used. This derives "natural" clusters by establishing disjoint density surfaces according to a probabilistic model (Wishart, 1978). Squared Euclidean distance was used to calculate the similarity matrix. (The standard CLUSTAN input parameters were invoked.)

Other graphical representations besides the dendrograms from the hierarchical methods were produced. For these, a principal components analysis was performed. Five factors were computed and filed. Then, two scatter diagrams were plotted, the first utilizing the first and second component factors for the axes, and the second utilizing the third and second component factors for axes. Also, graphic aids called cluster diagrams were plotted; these are outlines around each cluster's members. These two sets of two-dimensional representations are helpful in forming a sense of the amount of overlap among clusters as well as beginning to "see" the clusters' structure. (These were performed after it had been decided that four was the most appropriate number of clusters.)

To decide upon the most appropriate number of clusters in Sample A-R, numerous aspects of the collection of solutions were considered. The dendrograms were visually inspected to form an initial estimate of the number of clusters. The clusters that appeared in one solution were searched for in other solutions. If there are four clusters in the "true" structure, one might expect to find four clusters in most clustering solutions. Therefore, consistency of the number of clusters across solutions contributed to the decision. The means and variances for each cluster of a solution were compared to provide information about within- and between-cluster structure for that particular solution. Also, finding consistency of membership of clusters across methods suggests clusters that should be counted. The iterative partitioning solutions were cross-compared. Their results and those from any additional methods were scanned for cluster memberships that differed from those produced by hierarchical techniques. Homogeneity of membership within a cluster, in terms of the individuals' BSQ diagnostic assignments, was helpful in considering the clinical meaningfulness of each cluster. This amalgam of information was used to decide the most appropriate number of clusters, as well as to select the "best" solution, for Sample A-R. Centroid vectors were calculated for each cluster of the best solution.

Sample B. The exact procedure that produced the best solution for Sample A was replicated with Sample B. This started with the removal of outliers to form Sample B-R. Sample B-R was cluster analyzed, then, using the minimum variance method with squared Euclidean distance. The same procedure used with Sample A-R to relocate individuals was applied to the four-cluster solution. This yielded one classification of Sample B-R.

Validation. The nearest-centroid evaluation technique (McIntyre & Blashfield, 1980) was followed. Accordingly, the members of Sample B-R were assigned to the nearest centroid vector of Sample A-R. This yielded a second classification of Sample B-R. Then, the agreement between these two classifications of Sample B-R was measured with the kappa statistic. This provides an index of the goodness of the cluster solution which can be used to address the issue of validity.

CHAPTER VI
RESULTS

Findings for the Total Sample (N=200)

The means, standard deviations, and variances for each of the nine temperament categories, as measured by the Behavioral Style Questionnaire (BSQ), are presented in Table 6-1. These were compared to those reported for the BSQ standardization sample (McDevitt & Carey, 1978). Using the F test for equality of variances, only the threshold category was found to be significantly different $F(349,199) = 1.29$, $p<.05$. Then, t tests were used to compare the means. The t test for differences between the means for groups with heterogeneous variances was not significant for the threshold category. The other eight categories were compared using t tests for groups with homogeneous variances. The means for six categories were significantly different from those for the BSQ. This study's sample was found to be described as significantly more active, $t(199) = -2.29$, $p<.025$; more arrhythmic, $t = -5.88$, $p<.001$; more slowly adaptable, $t = -8.33$, $p<.001$; less intense, $t = 4.33$, $p<.001$; more nonpersistent, $t = -5.00$, $p<.001$; and less distractible, $t = 2.00$, $p<.05$. The magnitude of the actual differences was small (less than .5 s.d.). No significant

Table 6-1. Means, standard deviations, and variances for temperament categories (N=200).

Category	Mean	Standard Deviation	Variance
Activity	3.72	.75	.56
Rhythmicity	2.95	.63	.40
Approach/withdrawal	3.06	.89	.80
Adaptability	2.80	.79	.62
Intensity	4.39	.64	.41
Mood	3.34	.70	.48
Persistence	3.02	.71	.51
Distractibility	3.81	.78	.61
Threshold	3.89	.53	.28

differences were found for approach/withdrawal, mood, and threshold.

An analysis of variance revealed no significant age-sex interactions across all nine categories. An analysis of covariance for age and sex revealed only two significant effects; there was a significant sex effect for activity, $F(1) = 4.99$, $p<.05$, with boys being rated as somewhat more active than girls, and there was a significant age effect for distractibility, $F(1) = 6.62$, $p<.01$ with younger children being rated as somewhat more distractible.

BSQ diagnostic clusters. All subjects were assigned to one of the five diagnostic clusters of the BSQ. The distribution is as follows: 16.5% difficult (n=33; 11 girls/22 boys); 14.5% intermediate high (n=20; 14/15); 8.0% slow to warm up (n=16; 7/9); 27% intermediate low (n=54; 29/25); and 34% easy (n=68; 31/37). These assignments for difficult, easy, and slow to warm up children yielded a coverage of 58.5%. The distributions for boys and girls did not differ significantly, $\chi^2(4) = 3.52$, $p>.10$. The distributions across diagnostic clusters for this sample and for the BSQ standardization sample did not differ significantly, $\chi^2(4) = 4.94$, $p>.10$.

BSQ instrument information. Although the focus of this study is not upon the construction of the instrument per se, the data do provide pertinent information.

Of the 200 questionnaires, 137 (68.5%) had every item answered. Item 22, "The child picks up the nuances or subtleties of parental explanations," was left blank more often than any other item (n=11). Items 19 and 25 were each left blank 7 times, 49 of the 100 items were answered in every instance. To be included questionnaires (200 of 208) had no more than two items left blank of the total set and had at least 80% of the items within a category completed. The degree of correlation between categories ranged from $\underline{r}$=.01 (between distractibility and rhythmicity) to $\underline{r}$=.67 (between mood and adaptability). Most of the remaining correlations were in the low range ($\underline{r}$<.20, 20 of the correlations) with the rest in the low-moderate range. Distractibility and rhythmicity were the categories that correlated least with other scales. Adaptability showed the greatest extent of intercorrelation.

Descriptive comments were occasionally written on the questionnaire by the rater. These mostly involved objections to the repetitive quality of some of the items. Several parents objected to the sexist language of Item 21: "The child had trouble leaving the mother the first three days when he/she entered school." On Item 22, the item most frequently left blank, many made question marks in the margin.

Dividing the total sample. The total sample was divided in a random fashion into two groups: Sample A (n=99)

was comprised of 49 girls and 50 boys; Sample B (n=101) had 42 girls and 59 boys. There were no significant differences with respect to sex for the two groups, $\chi^2(1) = 1.60$, $p>.21$. There was slightly more variation in terms of age in the second group, $F(100,98) = 1.66$, $p<.02$ but the means were not significantly different.

Clustering Sample A

Single linkage with squared Euclidean distance did result in chaining (see Figure 3-1). The last 10 individuals were chained on; these individuals' z scores were examined across all 9 temperament categories. The chained-on individuals were also the individuals who had two or more z scores greater than two standard deviations. These 10 subjects (5 boys and 5 girls) were removed from Sample A, forming Sample A-R (Revised) (n=99, 44 girls, 45 boys).

A variety of clustering techniques were applied to the raw data of Sample A-R (see Chapter V for a discussion of standardization).

Visual inspection of the dendrograms produced by average linkage with Pearson product-moment correlation (see Figure 6-1) and by minimum variance with squared Euclidean distance (see Figure 6-2) suggested there were probably three to five clusters in the data. Consideration of replication of clusters across methods, consistency of

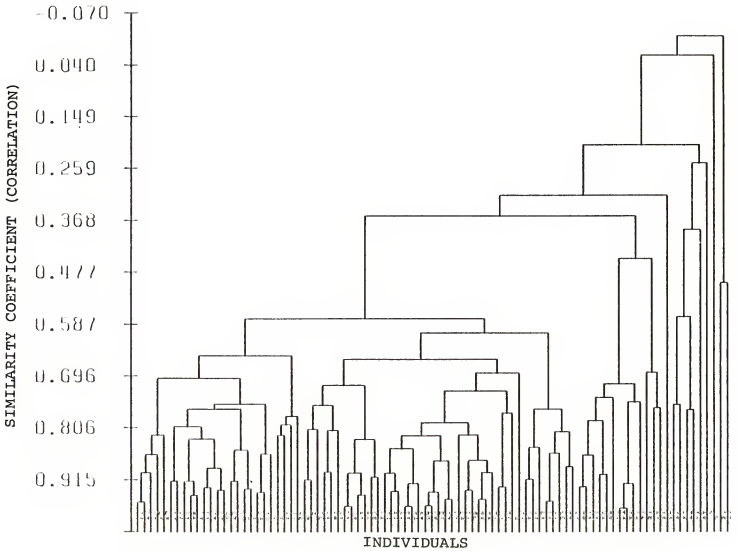

Figure 6-1. Average linkage with Pearson product-moment correlation, Sample A-R (n=89).

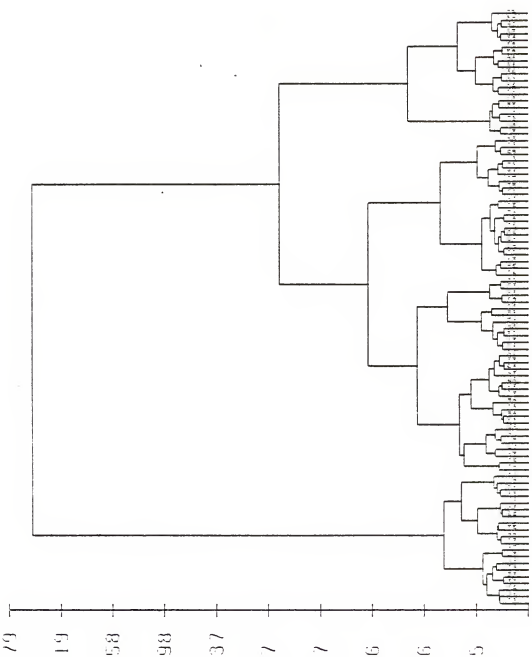

Figure 6-2. Minimum variance with squared Euclidean distance, Sample A-R (n=89).

membership, and meaningfulness strongly suggested the most appropriate number of clusters to be four. Consideration of the solutions obtained with iterative partitioning methods supported this decision, as did a comparison of the relative homogeneities of clusters in terms of members' BSQ diagnostic assignments for the four- and five-cluster solutions from the minimum variance method.

The BSQ diagnostic cluster assignments were used as an initial partition of the data. This method would "check" for poorly placed individuals. Even after the maximum number of iterations specified (10), a stable solution had not been reached for 5 clusters. A total of 76 relocations had occurred. After the fusion of the two most similar clusters, two iterations and nine relocations occurred before a stable solution was reached. A similar course was obtained when random assignments to 10 groups served as the initial partition. The four-cluster solution seemed best. The four-cluster solutions from these partitioning methods yielded fairly similar clusters to those in the four-cluster solution from the minimum variance method.

An iterative partitioning method (k-means) with error sum of squares was also used with the four-cluster and five-cluster solutions from the minimum variance method as initial partitions to "see" if relocating any "misfits" improved the solutions. For the 4-cluster initial grouping, 2 iterations were necessary before clusters were stable,

Table 6-2. The cluster means and sizes for the "best" solution (minimum variance with distance, followed by k-means procedure) for Sample A-R (n=89).

Cluster	Temperament Category								
	Activ	Rhy	App	Adap	Int	Mood	Pers	Dist	Thresh
1 n=24	3.40	2.60	2.35	1.86	4.10	2.66	2.40	3.93	3.76
2 n=24	3.71	2.90	2.91	2.73	4.78	3.57	3.04	4.43	4.20
3 n=19	4.44	3.30	3.88	3.74	4.85	4.14	3.78	3.96	4.11
4 n=22	3.58	3.12	3.41	2.92	4.14	3.23	3.10	3.33	3.90

Table 6-3. Composition of the four clusters considered the "best" solution for Sample A-R in terms of members' BSQ diagnostic cluster assignment.

Cluster	BSQ diagnostic assignment (% of cluster size)				
	Easy	Difficult	Slow to warm up	Interm. high	Interm. low
1	83.3	0	0	0	16.7
2	4.2	8.3	8.3	33.3	45.8
3	0	57.9	5.3	26.3	10.5
4	22.7	9.1	13.6	9.1	45.5

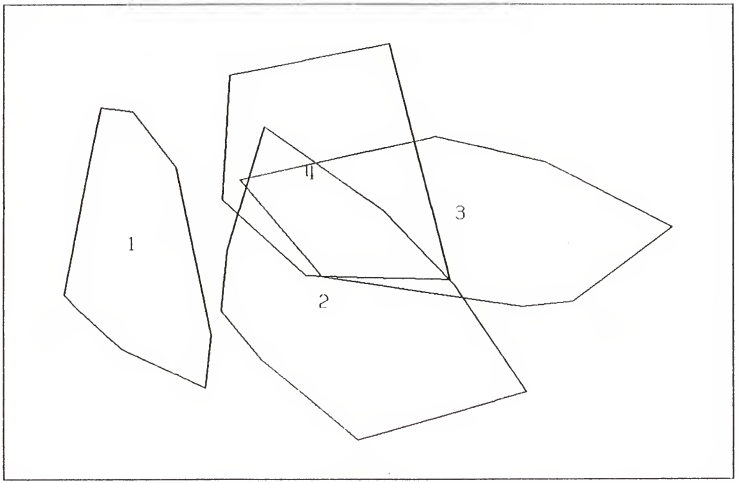

Figure 6-3. Outlines of clusters in principal components space, first by second factors, Sample A-R (n=89).

Figure 6-4. Outlines of clusters in principal components space, second by third factors, Sample A-R (n=89).

involving 17 relocations. This solution was selected from among all the solutions as the "best." Its procedure had been 1) clustering Sample A-R using the minimum variance method with squared Euclidean distance, and 2) relocating any poorly placed individuals in the four-cluster solution, using the k-means method of iterative partitioning with error sum of squares. Each cluster's size and its means across the temperament categories are given in Table 6-2. Consideration of the relationship of these clusters and the diagnostic clusters for each cluster's members was helpful in interpreting the clusters. Table 6-3 shows which diagnostic assignments made up each cluster. Graphic representations (see Figures 6-3 and 6-4) also aided conceptualization of the data.

The Nearest-Centroid Technique

Sample B was subjected to the same procedure that had been used to arrive at the solution considered the "best" from Sample A. This procedure started by performing a single linkage analysis of Sample B with squared Euclidean distance (see Figure 6-5) to assist in the identification of outliers. After checking these individuals' z scores, 8 were removed, forming Sample B-R (Revised) (n=93), of which 2 were girls, leaving 40, and 6 were boys, leaving 53. Sample B-R was then cluster-analyzed using minimum

Figure 6-5. Single linkage with squared Euclidean distance, Sample B (n=101).

variance with squared Euclidean distance (see Figure 6-6), followed by relocating any poorly located individuals using the k-means iterative partitioning method until stable for four clusters. The results are presented in Table 6-4. For purposes of comparison of the solutions from Sample A-R and Sample B-R, the graphic representation of the cluster outlines plotted on the first-by-second principal components axes is presented in Figure 6-7.

Agreement kappa. According to the nearest-centroid evaluation technique (McIntyre & Blashfield, 1980), the individuals of Sample B-R were assigned to the nearest centroid vector from Sample A-R. The agreement between the two classifications of Sample B-R was significant, $\underline{k}=.05$, $p<.01$. Thus, the extent of agreement is significantly greater than chance level; yet, given that kappa ranges from 0 (chance agreement) to 1 (perfect agreement), the magnitude of agreement is low.

Sample A-R and Sample B-R were compared in terms of the distribution across BSQ diagnostic cluster assignments. Sample A-R was not found to differ significantly from Sample B-R, $\chi^2(4) = 5.41$, $\underline{p}>.20$.

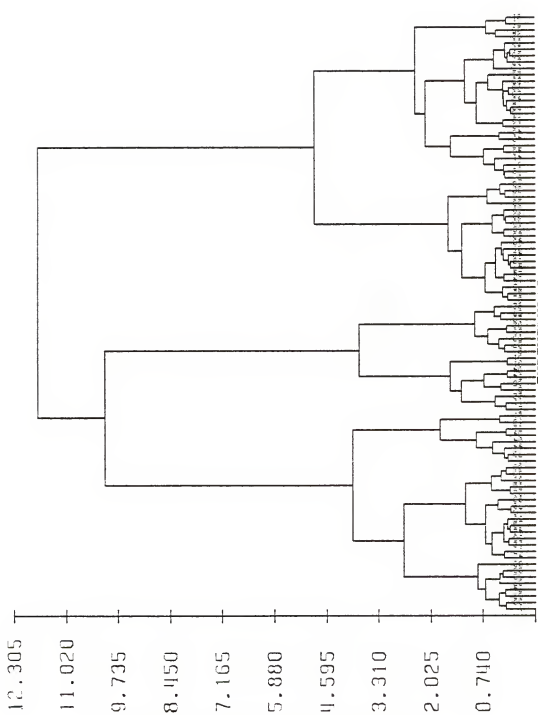

Figure 6-6. Minimum variance with squared Euclidean distance, Sample B-R (n=93).

Table 6-4. The cluster means and sizes for the "best" solution, as determined for Sample A-R, applied to Sample B-R (n=93).

Cluster	Temperament Category								
	Activ	Rhy	App	Adap	Int	Mood	Pers	Dist	Thresh
1 n=26	4.33	3.24	3.80	3.14	4.12	3.57	2.90	3.53	3.64
2 n=21	4.51	3.07	2.57	3.01	4.67	3.74	3.59	4.24	4.07
3 n=22	3.62	2.88	2.38	2.34	4.42	2.82	2.74	4.29	4.01
4 n=24	3.02	2.56	2.76	2.28	3.91	2.84	2.78	3.10	3.51

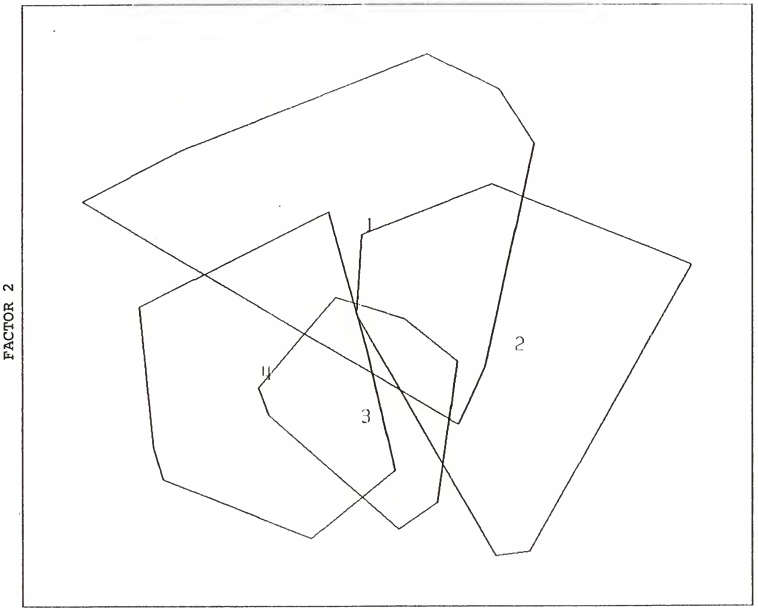

Figure 6-7. Outline of clusters in principal components space, first by second factors, Sample B-R (n=93).

CHAPTER VII
DISCUSSION

This study was designed to address two principal questions. The first question asked whether there are naturally existing groups of children, distinguishable on the basis of their temperament. Children's temperament was measured with the Behavioral Style Questionnaire (BSQ) (McDevitt & Carey, 1978), a parent rating instrument developed to measure Thomas and others' (1963, 1968, 1970, 1977) nine dimensions of temperament. The data were cluster analyzed using a cross-validation paradigm in order to establish the validity of the empirically derived solution. After some support had been established for the validity of the cluster solution, the study's second question was germaine. This asked whether the empirical types (i.e., the clusters as identified in the "best" cluster solution) are similar to the ideal types: the difficult child, the easy child, and the slow to warm up child (Thomas et al., 1968).

In the first phase of the cluster analysis, many methods were applied to half of the data, and a "best" solution was selected. It was produced by using the minimum variance method with squared Euclidean distance, followed by the k-means partitioning method to relocate any poorly assigned individuals. Finding the minimum variance method

to be superior to other methods is consistent with Milligan's (1981b) review of clustering methods. Unlike Milligan's conclusions, the average linkage method was not nearly as useful in initiating clusters. This may be a reflection of the amount of overlap among the clusters. It has been observed in Monte Carlo studies (Bayne, Beauchamp, Begovich, & Kane, 1980; Milligan, 1981b) that the minimum variance technique is generally superior to the group average technique when overlap is increased. In comparing similarity coefficients, it is interesting that correlation was relatively ineffective in initiating distinct clusters in comparison with distance. This suggests that the degree of elevation of category scores may be more useful than the shape of the profiles for distinguishing groups of children on the basis of their temperament (Skinner, 1978), at least as temperament is measured in this study.

The degree of stability of the "best" solution was found to be greater than what would be expected on the basis of chance alone, yet it was far from a strong showing. There are various possible explanations for the finding of low stability. It is possible that it has to do with the nearest-centroid evaluation procedure itself. McIntyre and Blashfield (1980) found that agreement kappa covaried with the degree of overlap among clusters. When there was a relatively large amount of overlap, agreement kappa generally was low. The four empirical temperament clusters do

overlap considerably. It may be, then, that this low estimate of the stability of the solution is more reflective of the evaluation method than the "true" stability. Also, although the effect is not understood, McIntyre and Blashfield have also found that this procedure can be adversely affected by the degree of intercorrelation among the variables. It may be that the lack of independence of the nine temperament dimensions acted to lower the estimate of stability, which estimates the solution's accuracy.

Alternatively, it is also possible that the identified empirical types are not particularly stable. The possible explanations for this low stability include 1) Sample A and Sample B might have been different, 2) the clustering techniques might have been differentially affected by some aspect of the data, 3) the selected solution was not actually a good one, i.e., not reflective of the "true" structure, and 4) the solution was well-chosen but reflective of a "true" structure that may not include particularly strong clear types.

Regarding the first explanation, given the random division of the total pool into two samples, it is improbable that the low stability is a reflection of significantly different structures for the separate samples. The finding of no significant differences in the samples' distributions across BSQ diagnostic cluster assignment corroborates such a conclusion. Regarding the second explanation, there was

slightly more variation in age in the second sample. Although age did not relate with eight of the nine temperament dimensions, suggesting that it is unlikely that this could account for the low degree of stability, it could be that the clustering methods were reactive to the variability. Further development in clustering methodology and theory as well as the construct of temperament is necessary to address the remaining explanations.

However, finding a significant degree of stability, even though small, provides some evidence for the accuracy, and, therefore, the validity, of the solution. The study's answer to the first question could, therefore, be answered that, yes, there may be temperament types.

To respond to the study's second question, the four clusters from Sample A-R were compared with those from Sample B-R. Three of the four clusters were nearly identical across samples. The fourth was so dissimilar that it is not considered a valid empirical type.

Interpretation of the three stable clusters progressed in stages. The characteristic profile for each cluster was considered first. The dimensions which correspond to the definitions of the ideal types (Thomas et al., 1968) were compared across solution. These were also compared with the operational definitions for the difficult, easy, slow to warm up, and intermediate groups (McDevitt & Carey, 1978). Then, the remainder of the nine dimensions were compared.

Next, to see whether the cluster description was actually representative of most of its members, the profiles for every individual were compared to the profile of its parent cluster.

It was found that the three stable clusters were quite consistent in their interpretability. Of the three, two were quite similar to one another. The children in these two groups were characterized by their regularity, positive approach responses to new situations and stimuli, high adaptability to change, and mild or moderately intense mood which is typically positive. This set of temperament characteristics matches that of the easy child as described by Thomas et al. and as operationalized by McDevitt and Carey. Three of the remaining four characteristics reveal the differences between these two empirically derived clusters. In order of degree (with greatest difference first), one group was much less distractible, had a higher sensory threshold, and was less active than the other group. The individual profiles are highly consistent with the parent profile. These clusters have been termed the nondistractible easy child and the distractible easy child. A consideration of the composition of these clusters in terms of the BSQ diagnostic assignments of their members corroborated these interpretations, especially when the diagnostic assignments were adjusted for the differences in the means between this sample and the BSQ standardization sample.

Finding two distinct empirical types of easy children was not anticipated. These two clusters accounted for 51.6% of the children (averaged across the two solutions). This is a larger proportion of the distribution than the 40% reported by Thomas et al. or the 34.3% (clinical scoring method) found by McDevitt and Carey. Although this comparison could be interpreted as suggestive of this sample having a greater proportion of easy children than found in either the NYLS or the BSQ, there were no significant differences between this sample and the BSQ sample in terms of their distributions across diagnostic clusters. It suggests that the two empirical types involve something more than a mere division of the easy children as previously defined, although this remains to be replicated. It also suggests the utility and importance of considering numerous dimensions in future research.

The distractible easy children's characteristics appear to have some degree of overlap with those for the group of children identified as having attention deficit disorders. Although the relationship between similar problems and temperament has appeared in the literature already (e.g., Carey, McDevitt, & Baker, 1979; Lambert & Windmiller, 1977), this finding is exciting for the direction it suggests for future research possibilities.

The third stable empirically derived cluster is characterized by somewhat irregular biological functions,

withdrawal from new stimuli followed by slow adaptability after repeated contact, moderate-to-mild intensity of reactions (whether positive or negative), and moderate-to-negative mood. These characteristics match the description of the slow to warm up child (Thomas et al., 1968). Although McDevitt and Carey's criteria vary somewhat from the Thomas et al. description, this empirical cluster also fits their operationalization. Additionally, this cluster is characterized by being less active than average, average in terms of persistence and sensory threshold, and non-distractible. Consideration of the fit of the BSQ assignments for each individual with the cluster profile also corroborated this description, especially when adjusted for differences between the groups' means. This empirical type has been termed the slow to warm up child. It comprised 26% of the sample (averaged across solutions) which, again, is greater than the proportions for either the NYLS or the BSQ samples (10% and 5.7%, respectively. Given that the overall distributions for the samples did not differ, the meaning of finding the rate to be higher is unclear.

The fourth cluster was not stable across samples and cannot be considered valid. In Sample A-R, it matched the description of the difficult child (Thomas et al.). However, the fourth cluster in Sample B-R was extremely dissimilar and not easily interpreted.

These findings raise a third question--what about the ideal type known as the difficult child? There are numerous factors which contribute to taxonomic variance (Achenbach, 1981). This study's lack of support for the validity of the difficult child may involve three possible explanations (Skinner, 1981): 1) the empirical typology does not measure the difficult child construct, 2) the theoretical definition is faulty or incorrect, and 3) this study failed to provide an adequate test.

It is possible that clustering methods failed to measure the difficult child type. This might have happened for two reasons. First, the actual proportion of difficult children in the sample might have been disproportionately small relative to the other clusters, resulting in the methods not picking out these children. The difficult children comprise the smallest group in both the NYLS and the BSQ. Second, it may be that the difficult children were so different from the rest of the sample that they appeared as outliers and were subsequently removed from the sample, prior to clustering. This explanation receives partial support following a review of the BSQ diagnostic assignments of the outliers. It was found that 7 of the total 18 outliers had been assigned to the difficult diagnostic cluster. This is 38.9% which is higher than what would be expected knowing that difficult children comprised only 16.5% of the total sample, 10% of the NYLS sample, and

12.6% of the BSQ standardization sample. It may be, then, that this empirical typology misses the difficult type. Future research involved in the empirical basis for a difficult type might benefit from clustering a sample chosen for its high proportion of clinically difficult children.

It is also possible that some bias was operative in the selection of this sample, not operative in previous work, that affected the adequacy of this "test." Although this sample was selected to be similar to that for the BSQ, the selection procedure could have introduced a bias. Although comparisons of the distributions across clusters revealed no significant differences for either type of facility or time of year for data collection, these variables may still have some effect. Also, the return rate for this sample was lower than for the BSQ (79.1% versus 94.9%). It is possible that parents of certain types of children are less likely to provide information about their children which could bias the results. Also, both the NYLS and the BSQ relied on subjects from private medical practices. This study recruited subjects from child care facilities. It is possible that this selection factor introduces bias. Perhaps difficult children are less likely to be enrolled in schools than other types of children. In other words, the base rates may vary across settings. The methodology presented in this study should be replicated across settings.

In sum, the empirical typology derived in this study offers some support for the validity of two of Thomas and others' (1968) ideal types--the easy child and the slow to warm up child. The empirical typology is limited in terms of its internal properties: the robustness across samples was low, although statistically significant; the clusters were somewhat homogeneous within clusters but the degree of overlap indicates insufficient heterogeneity between clusters; and the coverage is moderate. It is clear that further work is necessary for the development of a sound empirical typology of temperament.

The lack of support for the difficult child ideal type is not presented as indicative that such a type does not exist. It should be remembered that cluster analysis is a descriptive tool for discovery. All that can be stated on the basis of these findings is that the method did not "find" any difficult type for this sample.

Implications for future research are several. First, it seems that the measurement of temperament needs refinement. The finding of significant differences between the means of this group and those for the BSQ standardization sample raises the question of the generalizability of that instrument. Also, the replication of finding the dimensions to be intercorrelated points to the necessity of further work to develop independent dimensions of temperament. These suggestions are consistent with a recent review of

temperament measurement (Hubert, Wachs, Peters-Martin, & Gandour, 1982). Also, inasmuch as this study worked with perceived temperament, further work is necessary to determine its relevance to actual characteristics of the child.

The findings of two distinct groups of easy children points to the importance of considering many dimensions of temperament in order to establish a good classification. The meaning of this finding is unclear, although the differentiating characteristics are suggestive of an avenue for future exploration. Certainly, the low stability of the empirical typology and the questions it raises, especially concerning the validity of the difficult child ideal type, confirm the opinion that there is an inadequate empirical basis at present to justify the applied use of temperament classification.

The developmental information is also interesting. Only one sex difference was observed. Finding boys more active than girls is consistent with the findings of McDevitt and Carey (1978). Finding so few sex effects contradicts the conclusions of Maurer et al. (1980). Only one age effect was observed. Finding younger children to be more distractible is consistent with the findings of McDevitt and Carey. Although the implications of these differences are unknown at present, more methodological and theoretical work will lead to a better understanding

of the nature of children's temperament and its role in their development.

REFERENCE NOTES

1. Bates, J.E. Temperament as a part of social relationships: Implications of perceived infant difficultness. Paper presented at the International Conference on Infant Studies, Austin, Texas, March 1982.

2. Milligan, G.W., Soon, S.C., & Sokol, L.M. The effect of cluster size, dimensionality, and the number of clusters on recovery of true cluster structure (Working Paper Series 81-71). Unpublished manuscript, Ohio State University, 1981.

3. Keogh, B.K., Kornblau, B.S., & Ballard-Campbell, M. Summary report: Temperament and attribution studies, Project REACH, 1977-1982. Unpublished manuscript, University of California at Los Angeles, 1982.

REFERENCES

Achenbach, T.J. The role of taxonomy in developmental psychopathology. In M.E. Lamb & A.L. Broun (Eds.), Advances in Developmental Psychology (Vol. 1). Hillsdale, N.J.: Erlbaum, 1981.

Achenbach, T.M., & Edelbrock, C.S. The classification of child psychopathology: A review and analysis of empirical efforts. Psychological Bulletin, 1978, 85, 1275-1301.

Aldenderfer, M.S., & Blashfield, R.K. Computer programs for performing hierarchical cluster analysis. Applied Psychological Measurement, 1978, 2(3), 403-411.

American Psychological Association. Standards for Educational and Psychological Tests. Washington, D.C.: Author, 1974.

Anderberg, M.R. Cluster Analysis for Applications. New York: Academic Press, 1973.

Bailey, K.D. Monothetic and polythetic typologies and their relation to conceptualization, measurement and scaling. American Sociological Review, 1973, 38, 13-88.

Bartko, J.J., Strauss, J.S., & Carpenter, W.T. An evaluation of taxonomic techniques for psychiatric data. Classification Society Bulletin, 1971, 2, 2-28.

Bates, J.E. The concept of difficult temperament. Merrill-Palmer Quarterly, 1980, 26(4), 299-319.

Bates, J.E., Olson, S.L., Pettit, G.S., & Bayles, K. Dimensions of individuality in the mother-infant relationship at six months of age. Child Development, 1982, 53, 446-461.

Bayne, C.K., Beauchamp, J.J., Begovich, C.L., & Kane, V.E. Monte Carlo comparisons of selected clustering procedures. Pattern Recognition, 1980, 12, 51-62.

Beebe, B., & Sloate, P. Assessment and treatment of difficulties in mother-infant attunement in the first three years of life: A case history. In J.D. Call, E.

Galenson, & R.L. Tyson (Eds.), Frontiers of Infant Psychiatry. New York: Basic Books, Inc., 1983.

Berberian, K.E., & Snyder, S.S. The relationship of temperament and stranger reaction for younger and older infants. Merrill-Palmer Quarterly, 1982, 28, 79-94.

Billman, J., & McDevitt, S.C. Convergence of parent and observer ratings of temperament with observations of peer interaction in nursery school. Child Development, 1980, 51, 395-400.

Blashfield, R.K. On the equivalence of four software programs for performing hierarchical cluster analysis. Psychometrika, 1977, 42, 429-431.

Blashfield, R.K. The growth of cluster analysis: Tryon, Ward, and Johnson. Multivariate Behavioral Research, 1980a, 15, 439-458.

Blashfield, R.K. Propositions regarding the use of cluster analysis in clinical research. Journal of Consulting and Clinical Psychology, 1980, 48(4), 456-459.

Blashfield, R.K., & Aldenderfer, M.S. Computer programs for performing iterative partitioning cluster analysis. Applied Psychological Measurement, 1978a, 2(4), 533-541.

Blashfield, R.K., & Aldenderfer, M.S. The literature on cluster analysis. Multivariate Behavioral Research, 1978b, 13, 271-295.

Blashfield, R.K., & Draguns, J.G. Evaluative criteria for psychiatric classification. Journal of Abnormal Psychology, 1976, 85, 140-150.

Braun, C. Teacher expectations: Sociopsychological dynamics. Review of Educational Research, 1976, 46, 185-213.

Brazelton, T.B. Infants and Mothers: Differences in Development. New York: Dell Publishing Company, 1969.

Brazelton, T.B. Neonatal Behavioral Assessment Scale. Philadelphia: J.B. Lippincott, 1973.

Brazelton, T.B. Assessment techniques for enhancing infant development. In J.D. Call, E. Galenson, & R.L. Tyson (Eds.), Frontiers of Infant Psychiatry. New York: Basic Books, Inc., 1983.

Buss, A.H., & Plomin, R. A Temperament Theory of Personality Development. New York: John Wiley, 1975.

Cadoret, R.J., Cunningham, L., Loftus, R., & Edwards, J.E. Studies of adoptees from psychiatrically disturbed biologic parents: II. Temperamental, hyperactive, antisocial, and developmental variables. Journal of Pediatrics, 1975, 87, 301-306.

Cameron, J. Parental treatment, children's temperament, and the risk of childhood behavior problems: 1. Relationships between parental characteristics and changes in children's temperament over time. American Journal of Orthopsychiatry, 1977, 47(4), 568-576.

Cameron, J. Parental treatment, children's temperament, and the risk of childhood behavioral problems: 2. Initial temperament, parental attitudes, and the incidence and form of behavioral problems. American Journal of Orthopsychiatry, 1978, 48(1), 140-147.

Campbell, S.B.G. Mother-infant interaction as a function of maternal ratings of temperament. Child Psychiatry and Human Development, 1979, 10(2), 67-76.

Carey, W.B. A simplified method for measuring infant temperament. Journal of Pediatrics, 1970, 77, 188-194.

Carey, W.B. Clinical applications of infant temperament measurements. Journal of Pediatrics, 1972a, 81, 823-828.

Carey, W.B. Measuring infant temperament. Journal of Pediatrics, 1972b, 81, 414.

Carey, W.B. Night waking and temperament in infancy. Journal of Pediatrics, 1974, 84, 756-758.

Carey, W.B., Fox, M., & McDevitt, S.C. Temperament as a factor in early school adjustment. Pediatrics, 1977, 60(4), 621-624.

Carey, W.B., & McDevitt, S.C. A revision of the infant temperament questionnaire. Pediatrics, 1978a, 61, 735-738.

Carey, W.B., & McDevitt, S.C. Stability and change in individual temperament diagnoses from infancy to early childhood. Journal of the American Academy of Child Psychiatry, 1978b, 17, 331-338.

Carey, W.B., & McDevitt, S.C. Commentary: Measuring infant temperament. Journal of Pediatrics, 1980, 96(3 Pt. 1), 423-425. (Letter)

Carey, W.B., McDevitt, S.C., & Baker, D. Differentiating minimal brain dysfunction and temperament. Developmental Medicine and Child Neurology, 1979, 21(6), 765-772.

Carroll, R.M., & Field, J. A comparison of the classification accuracy of profile similarity measures. Multivariate Behavioral Research, 1974, 9, 373-380.

Cattell, R.B. Personality: A Systematic and Factual Study. New York: McGraw-Hill, 1950.

Cattell, R.B. The three basic factor-analytic research designs--their inter-relations and derivatives. Psychological Bulletin, 1952, 49, 499-502.

Cattell, R.B., & Coulter, M.A. Principles of behavioral taxonomy and the mathematical basis of the taxonome computer program. The British Journal of Mathematical and Statistical Psychology, 1966, 19(2), 237-269.

Chamberlin, R.W. Relationships between child-rearing styles and child behavior over time. American Journal of Disease in Children, 1978, 132, 155-160.

Chess, S., & Hassibi, M. Behavior derivations in mentally retarded children. Journal of the American Academy of Child Psychiatry, 1970, 9, 282-297.

Chess, S., & Korn, S. Temperament and behavior disorders in mentally retarded children. Archives of General Psychiatry, 1970, 23, 122-130.

Chess, S., Korn, S., & Fernandez, P. Psychiatric Disorders of Children with Congenital Rubella. New York: Brunner/Mazel, 1971.

Cohen, J. A coefficient of agreement for nominal scales. Educational and Psychological Measurement, 1960, 20, 37-46.

Cormack, R.M. A review of classification. The Journal of the Royal Statistical Society (Series A), 1971, 134, 321-367.

Cronbach, L.J., & Gleser, G.C. Assessing similarity between profiles. Psychological Bulletin, 1953, 50, 456-473.

Cronbach, L.J., & Meehl, P.E. Construct validity in psychological tests. Psychological Bulletin, 1955, 52 (4), 281-302.

Dreger, R.M. General temperament and personality factors related to intellectual performances. Journal of Genetic Psychology, 1968, 113, 275-293.

Driver, H.W., & Kroeber, A.L. Quantitative expression of cultural relationships. University of California Publications in Archaeology and Ethnology, 1932, 31, 211-256.

Dubes, R., & Jain, A.K. Validity studies in clustering methodologies. Pattern Recognition, 1979, 11, 235-254.

Dunn, J., & Kendrick, C. Studying temperament and parent-child interaction: Comparison of interview and direct observation. Developmental Medicine and Child Neurology, 1980, 22, 484-496.

Edelbrock, C. Mixture model tests of hierarchical clustering algorithms: The problem of classifying everybody. Multivariate Behavioral Research, 1979, 14, 367-384.

Egeland, B., & Deinard, A. Psychometric and theoretical credibility of three measures of infant temperament. Research Relating to Children, 1978, Bull. 41, 41-EA-1.

Everitt, B. Cluster Analysis. London: Heinemann Educational Books Ltd., 1974.

Everitt, B.S. Unresolved problems in cluster analysis. Biometrics, 1979, 35, 169-181.

Fleiss, J.L., Cohen, J., & Everitt, B.S. Large sample standard errors of kappa and weighted kappa. Psychological Bulletin, 1969, 72, 323-327.

Fleiss, J.L., & Zubin, J. On the methods and theory of clustering. Multivariate Behavioral Research, 1969, 4, 235-250.

Garside, R.F., Birch, H., Scott, D. McI., Chambers, S., Kolvin, I., Tweddle, E.G., & Barber, L.M. Dimensions

of temperament in infant school children. Journal of Child Psychology and Psychiatry, 1975, 16, 219-231.

Garside, R.E., & Roth, M. Multivariate statistical methods and problems of classification in psychiatry. British Journal of Psychiatry, 1978, 133, 53-67.

Gordon, E.M., & Thomas, A. Children's behavioral style and the teacher's appraisal of their intelligence. Journal of School Psychology, 1967, 5, 292-300.

Graham, P., Rutter, M., & George, S. Temperamental characteristics as predictors of behavioral disorders in children. American Journal of Orthopsychiatry, 1973, 43, 328-339.

Gunn, P., Berry, P., & Andrews, R.J. The temperament of Down's syndrome infants: A research note. Journal of Child Psychology and Psychiatry, 1981, 22, 189-194.

Hempel, C.G. Aspects of Scientific Explanation. New York: Free Press, 1966.

Hertzig, M.E. Neurologic findings in prematurely born children at school age. In D. Ricks, A. Thomas, & M. Roff (Eds.), Life History Research in Psychopathology (Vol. III). Minneapolis: University of Minnesota Press, 1974.

Hertzig, M.E., Birch, H.G., Thomas, A., & Mendez, O.A. Class and ethnic differences in the responsiveness of preschool children to cognitive demands. Monographs of the Society for Research in Child Development, 1968, 33, 1-69.

Hubert, L. Kappa revisited. Psychological Bulletin, 1977, 84, 289-297.

Hubert, N.C., Wachs, T.D., Peters-Martin, P., & Gandour, M.J. The study of early temperament: Measurement and conceptual issues. Child Development, 1982, 53, 571-600.

Jardine, N., & Sibson, R. The construction of hierarchic and non-hierarchic classifications. Computer Journal, 1968, 11, 177-184.

Kagan, J. The construct of difficult temperament: A reply to Thomas, Chess, and Korn. Merrill-Palmer Quarterly, 1982, 28, 21-24.

Keogh, B.K. Children's temperament and teachers' decisions. In M. Rutter, R. Porter, & G.M. Collins (Eds.), Temperamental Differences in Infants and Young Children. London: Putnam Books Ltd., 1982a.

Keogh, B.K. Temperament: An individual difference of importance in intervention programs. Topics in Early Childhood Special Education, 1982b, 2, 25-31.

Keogh, B.K., & Pullis, M.E. Temperamental influences on the development of exceptional children. In B.K. Keogh (Ed.), Advances in Special Education (Vol. 1). Greenwich, Conn.: JAI Press, Inc., 1980.

Keogh, B.K., Pullis, M.E., & Cadwell, J. A short form of the Teacher Temperament Questionnaire. Journal of Educational Measurement, 1982, 19, 323-329.

Korner, A.F. Individual differences in neonatal activity: Implications for the origins of different coping styles. In J.D. Call, E. Galenson, & R.L. Tyson (Eds.), Frontiers of Infant Psychiatry. New York: Basic Books, Inc., 1982.

Lambert, N.M., & Windmiller, M. An exploratory study of temperament traits in a population of children at risk. Journal of Special Education, 1977, 11, 37-47.

Lerner, R.M., Palermo, M., Spiro III, A., & Nesselrode, J.R. Assessing the dimensions of temperamental individuality across the life span: The dimensions of temperament survey. Child Development, 1982, 53, 149-159.

Lyon, M.E., & Plomin, R. The measurement of temperament using parental ratings. Journal of Child Psychology and Psychiatry and Allied Disciplines, 1981, 22, 47-53.

Marcus, J., Thomas, A., & Chess, S. Behavioral individuality in kibbutz children. The Israel Annals of Psychiatry and Related Disciplines, 1969, 7(1), 43-54.

Matheny, A.P., Jr., & Dolan, A.B. A twin study of personality and temperament during middle childhood. Journal of Research in Personality, 1980, 14, 224-233.

Maurer, R., Cadoret, R.J., & Cain, C. Cluster analysis of childhood temperament data on adoptees. American Journal of Orthopsychiatry, 1980, 50, 522-534.

McDevitt, S.C., & Carey, W.B. The measurement of temperament in three- to seven-year-old children. Journal of Child Psychology and Psychiatry, 1978, 19. 245-253.

McInerny, T., & Chamberlin, R.W. Is it feasible to identify infants who are at risk for later behavioral problems? The Carey temperament questionnaire as a prognostic tool. Clinical Pediatrics, 1978, 17, 233-238.

McIntyre, R.M., & Blashfield, R.K. A nearest-centroid technique for evaluating the minimum-variance clustering procedure. Multivariate Behavioral Research, 1980, 15, 225-238.

McQuitty, L.L. A mutual development of some typological theories and pattern-analytic methods. Educational and Psychological Measurement, 1967, 27, 21-46.

Milligan, G.W. A Monte Carlo study of thirty internal criterion measures for cluster analysis. Psychometrika, 1981a, 46, 187-199.

Milligan, G.W. A review of Monte Carlo tests of cluster analysis. Multivariate Behavioral Research, 1981b, 16, 379-407.

Persson-Blennow, I., & McNeil, T.F. A questionnaire for measurement of temperament in six-month-old infants: Development and standardization. Journal of Child Psychology and Psychiatry, 1979, 20(1), 1-13.

Persson-Blennow, I., & McNeil, T.F. Temperament characteristics of children in relation to gender, birth order, and social class. American Journal of Orthopsychiatry, 1981, 51(4), 710-714.

Plomin, R. Behavioural genetics and temperament. In R. Porter, & G. Lawrenson (Eds.), Temperamental Differences in Infants and Young Children. London: Pitman Books Ltd., 1982a.

Plomin, R. The difficult concept of temperament: A response to Thomas, Chess, and Korn. Merrill-Palmer Quarterly, 1982b, 28, 25-33.

Pullis, M.E. An investigation of the relationship between children's temperament and school adjustment. Doctoral dissertation, University of California, Los Angeles, 1979.

Pullis, M.E., & Cadwell, J. The influence of children's temperament characteristics on teachers' decision strategies. American Journal of Educational Research, 1982, 19, 165-181.

Rosenhan, D.L. On being sane in insane places. Science, 1973, 179, 250-258.

Rothbart, M.K. The concept of difficult temperament: A critical analysis of Thomas, Chess, and Korn. Merrill-Palmer Quarterly, 1982, 28, 35-40.

Rothbart, M.K., & Derryberry, D. Development of individual differences in temperament. In M.E. Lamb, & A.L. Brown, Eds.), Advances in Developmental Psychology (Vol. 1). Hillsdale, New Jersey: Lawrence Erlbaum Associates, 1981.

Rowe, D.C., & Plomin, R. Temperament in early childhood. Journal of Personality Assessment, 1977, 41(2), 150-156.

Sameroff, A.J., Seifer, R., & Elias, P.K. Sociocultural variability in infant temperament ratings. Child Development, 1982, 53, 164-173.

Sawrey, W.L., Keller, L., & Conger, J.J. An objective method of grouping profiles by distance functions and its relation to factor analysis. Educational and Psychological Measurement, 1960, 20, 651-673.

Scholom, A., & Schiff, G. Relating infant temperament to learning disabilities. Journal of Abnormal Child Psychology, 1980, 8, 127-132.

Simonds, M.P., & Simonds, J.F. Relationship of maternal parenting behaviors to preschool children's temperament. Child Psychiatry and Human Development, 1981, 12, 19-31.

Skinner, H.A. Differentiating the contribution of elevation, scatter and shape in profile similarity. Educational and Psychological Measurement, 1978, 38, 297-308.

Skinner, H.A. Toward the integration of classification theory and methods. Journal of Abnormal Psychology, 1981, 90, 68-87.

Skinner, H.A., & Blashfield, R.K. Increasing the impact of cluster analysis research: The case of psychiatric classification. Journal of Consulting and Clinical Psychology, 1982, 50(5), 727-735.

Skinner, H.A., & Lei, H. Modal profile analysis: A computer program for classification research. Educational and Psychological Measurement, 1980, 40, 769-772.

Sokal, R.R. Classification: Purposes, principles, progress, prospects. Science, 1974, 185, 1115-1123.

Sokal, R.R., & Sneath, P.H.A. Principles of Numerical Taxonomy. San Francisco: W.H. Freeman, 1963.

Standley, K., Soule, A.B., Copans, S.A., & Klein, R.P. Multidimensional sources of infant temperament. Genetic Psychology Monographs, 1978, 98, 203-231.

Terestman, N. Mood quality and intensity in nursery school children as predictors of behavior disorder. American Journal of Orthopsychiatry, 1980, 50, 125-130.

Thomas, A., Birch, H.G., Chess, S., & Robbins, L.C. Individuality in responses of children to similar environmental situations. American Journal of Psychiatry, 1961, 117, 434-441.

Thomas, A., & Chess, S. An approach to the study of sources of individual differences in child behavior. Journal of Clinical Experimental Psychopathology and Quarterly Review in Psychiatric Neurology, 1957, 18, 347-357.

Thomas, A., & Chess, S. Temperament and Development. New York: Brunner/Mazel, 1977.

Thomas, A., & Chess, S. The Dynamics of Psychological Development. New York: Brunner/Mazel, 1980.

Thomas, A., Chess, S., & Birch, H.G. Temperament and Behavior Disorders in Children. New York: New York University Press, 1968.

Thomas, A., Chess, S., & Birch, H.G. The origin of personality. Scientific American, 1970, 223, 102-109.

Thomas, A., Chess, S., Birch, H.G., Hertzig, M.E., & Korn, S. Behavioral Individuality in Early Childhood. New York: New York University Press, 1963.

Thomas, A., Chess, S., & Korn, S.J. The reality of difficult temperament. Merrill-Palmer Quarterly, 1982, 28, 1-20.

Tryon, R.C., & Bailey, D.C. Cluster Analysis. New York: McGraw-Hill, 1970.

Tversky, A. Features of similarity. Psychological Review, 1977, 84, 327-352.

Ward, J.H. Hierarchical grouping to optimize an objective function. Journal of the American Statistical Association, 1963, 58, 236-244.

Wishart, D. Mode analysis. In A.J. Cole (Ed.), Numerical Taxonomy. New York: Academic Press, 1969.

Wishart, D. Clustan User Manual (3rd ed.). Edinburgh, Scotland: Edinburgh University, 1978.

Wood, A.L. Ideal and empirical typologies for research in deviance and control. Sociology and Social Research, 1969, 53, 227-241.

Yarrow, M.R. Problems of methods in parent-child research. Child Development, 1963, 34, 215-226.

BIOGRAPHICAL SKETCH

The author was born in Boston, Massachusetts, in 1953. She was raised in Needham, Massachusetts, where she attended public schools. She spent her senior year of high school in Sweden with a family while attending the gymnasiet. She attended Mount Holyoke College in South Hadley, Massachusetts, for her first year of college, after which she transferred to Brown University in Providence, Rhode Island. In 1975 she graduated magna cum laude with a Bachelor of Science degree in psychology. The following fall she entered graduate school, studying clinical psychology, at the University of Florida, Gainesville, Florida. During the first two years of graduate study, she was awarded a USPHS traineeship. Later, she was supported, in part, by research assistantships. She received clinical training at the University of Florida's Student Mental Health facility, and at the Psychology Clinic and the Children's Mental Health Unit, both at the University's teaching hospital. In 1980 she completed an internship in clinical psychology at Massachusetts General Hospital, Harvard Medical School, and Erich Lindemann Mental Health Center. Since then she has been a member of the faculty at Phillips Exeter Academy, Exeter, New Hampshire, working as a psychologist. Her primary professional interests include psychodynamic psychotherapy and

personality development within the family. In the future she plans to build a private practice and teach human development and personality to graduate and medical students.

I certify that I have read this study and that in my opinion it conforms to acceptable standards of scholarly presentation and is fully adequate, in scope and quality, as a dissertation for the degree of Doctor of Philosophy.

Jacquelin R. Goldman, Chairperson
Professor of Clinical Psychology

I certify that I have read this study and that in my opinion it conforms to acceptable standards of scholarly presentation and is fully adequate, in scope and quality, as a dissertation for the degree of Doctor of Philosophy.

Benjamin Barger
Professor of Clinical Psychology

I certify that I have read this study and that in my opinion it conforms to acceptable standards of scholarly presentation and is fully adequate, in scope and quality, as a dissertation for the degree of ~~Doctor of Philosophy~~.

W. Keith Berg
Associate Professor of Psychology

I certify that I have read this study and that in my opinion it conforms to acceptable standards of scholarly presentation and is fully adequate, in scope and quality, as a dissertation for the degree of Doctor of Philosophy.

Randy L. Carter
Assistant Professor of Biostatistics

I certify that I have read this study and that in my opinion it conforms to acceptable standards of scholarly presentation and is fully adequate, in scope and quality, as a dissertation for the degree of Doctor of Philosophy.

Everette E. Hall
Associate Professor of Clinical Psychology

This dissertation was submitted to the Graduate Faculty of the College of Health Related Professions and to the Graduate Council, and was accepted as partial fulfillment of the requirements for the degree of Doctor of Philosophy.

December 1983

Dean, College of Health Related Professions

Dean for Graduate Studies and Research